

What effect does short term Study Abroad (SA) have on learners'
vocabulary knowledge?

Thomas Caton

Submitted to Swansea University in fulfilment of the requirements
for the Degree of Doctor of Philosophy in Applied Linguistics
Swansea University

Swansea University
March 2023

Copyright: The Author, Thomas Caton, 2023.

Abstract

This thesis describes a study which tracks longitudinal changes in vocabulary knowledge during a short-term Study Abroad (SA) experience. A test of productive vocabulary knowledge, Lex30 (Meara & Fitzpatrick, 2000), requiring the production of word association responses, is used to elicit vocabulary from 38 Japanese L1 learners of English at four test times at equal intervals before and after an SA experience.

The study starts by investigating whether there are changes in both the total number of words and in the number of less frequently occurring words produced by SA participants. Three additional ways of measuring the development of lexical knowledge over time are then proposed. The first examines changes in the ability of participants of different proficiency levels in producing collocates in response to Lex30 cue words. The second tracks changes in spelling accuracy to measure if improvements take place over time. The third analysis uses an online measuring instrument (Wmatrix; Rayson, 2009) to explore if there are any changes in the mastery of specific semantic domains.

The results show that there is significant growth in the productive use of less frequent vocabulary knowledge during the SA period. There is also an increase in collocation production with lower proficiency participants and evidence of some improvement in the way certain vocabulary items are spelled. The tendency for SA learners to produce more words from semantic groups related to SA experiences is also demonstrated. Post-SA tests show that while some knowledge attrition occurs it does not decline to pre-SA levels. The study shows how short-term SA programmes can be evaluated using a word association test, contributing to a better understanding of how vocabulary develops during intensive language learning experiences. It also demonstrates the gradual shift of productive vocabulary knowledge from partial word knowledge to a more complete state of productive mastery.

Declaration and Statements

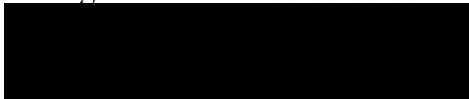
DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

 (candidate) Date *31st March 2023*


STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

 (candidate) Date *31st March 2023*


STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Swansea University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

 (candidate) Date *31st March 2023*

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

 (candidate) Date *31st March 2023*

Acknowledgements

A great many people have assisted me along the way in producing this thesis. First and foremost I should acknowledge my supervisor Tess Fitzpatrick, who got me started on this journey by introducing me to the Lex30 vocabulary task and highlighting numerous ways in which ways it could help with my research. Mention must also be made of Federica Barbiera for helping me with the final write up and Alison Wray for acquainting with the concept of semantic diversity.

Many friends and colleagues gave practical support in various forms. In particular I wish to thank Caroline Handley for informing me about the Glasgow norms, Andrew Wimhurst for giving me advice on the various ways Lex30 can be administered and Kiyoshi Yoneda for giving me valuable advice on the mysterious world of statistics.

I am very grateful to Chris Spzilman, Elena Téllez and John Druce who encouraged me to start this thesis and have maintained their strong interest and support in my progress throughout this long journey.

I also owe thanks to the many students who kindly gave up their time to provide me with the data on which this thesis is based.

Finally, I would like to thank my family, and in particular my wife, Tomoko Caton, who has provided endless support, both practical and moral.

Table of Contents

<i>Abstract</i>	II
<i>Declaration and Statements</i>	III
<i>Acknowledgements</i>	IV
<i>List of tables</i>	IX
<i>List of figures</i>	XI
<i>Abbreviations</i>	XII
Chapter 1: Introduction	1
1.1 <i>Evaluating Study Abroad programmes</i>	1
1.2 <i>Measuring vocabulary knowledge</i>	2
1.3 <i>Aspects of vocabulary knowledge</i>	3
1.4 <i>The organization of this thesis</i>	6
1.5 <i>The impact of COVID-19 on study abroad research</i>	7
Chapter 2: Literature review	9
2.1 <i>Study Abroad (SA) Definition and history</i>	9
2.1.1 <i>The history and growth of SA</i>	11
2.1.2 <i>The impact of the COVID-19 global pandemic</i>	13
2.2 <i>Japan and Study Abroad (SA)</i>	14
2.2.1 <i>Japanese Study Abroad (SA): Recent trends</i>	15
2.2.2 <i>Japanese Study Abroad (SA): Future challenges</i>	18
2.3 <i>Evaluating SA programmes</i>	21
2.3.1 <i>At Home (AH) versus Study Abroad (SA)</i>	24
2.3.2 <i>Alternatives to AH and SA designs</i>	25
2.4 <i>SA Research: Aspects of language change</i>	26
2.4.1 <i>Challenges in SA methodology</i>	27
2.4.2 <i>SA research and vocabulary change</i>	27
2.4.3 <i>Vocabulary knowledge and reading skills</i>	29
2.4.4 <i>The relationship between vocabulary knowledge and other skills</i>	31
2.5 <i>Research on SA vocabulary acquisition</i>	32
2.5.1 <i>Different aspects of vocabulary development</i>	32
2.5.2 <i>Some examples of vocabulary knowledge research</i>	33
2.6 <i>Measuring language change over short periods</i>	37
2.6.1 <i>New approaches to measuring change</i>	37
2.7 <i>Comparisons of productive vocabulary tests</i>	43
2.7.1 <i>Capturing vocabulary knowledge: Fitzpatrick and Clenton (2017)</i>	44
2.7.2 <i>Implications of Fitzpatrick and Clenton (2017)</i>	46
2.8 <i>Lex30 explanation and validation</i>	48
2.8.1 <i>Lex30 description</i>	49
2.8.2 <i>Lex30: Possible limitations</i>	52
2.8.3 <i>Reviews of Lex30</i>	54
2.9 <i>Conclusion</i>	57
Chapter 3: Lex30: a replication study	60
3.1 <i>The Lex30 productive vocabulary test</i>	61
3.2 <i>Measuring changes in vocabulary knowledge: Previous studies</i>	62

3.2.1	Fitzpatrick and Clenton (2010).....	62
3.2.2	Fitzpatrick’s 2003 Thesis	64
3.2.3	Fitzpatrick’s 2003 longitudinal study	65
3.2.4	Towards a replication experiment	66
3.3	<i>Methodology</i>	67
3.4	<i>Results</i>	69
3.5	<i>Discussion</i>	70
3.5.1	Participants	72
3.5.2	Participation period	72
3.5.3	Proficiency level.....	73
3.5.4	Timing of pre and post tests	73
3.5.5	Scoring protocols	74
3.5.6	Percentage score v. Raw score	75
3.6	<i>Conclusion</i>	76
Chapter 4: A closer look at Lex30 scores before and after SA		77
4.1	<i>Lex30 administration</i>	78
4.1.1	Scoring, word-bands and data analysis.	79
4.2	<i>Research Questions</i>	79
4.3	<i>Methodology</i>	80
4.3.1	Participants	80
4.3.2	SA programme	80
4.3.3	Test administration.....	81
4.3.4	Lex30 cue words and scoring protocol.....	82
4.3.5	Methods of analysis.....	86
4.4	<i>Results</i>	87
4.4.1	General overview	87
4.4.2	Changes in the total number of words produced (RQ1)	90
4.4.3	Changes in the number of infrequent words (RQ2).....	92
4.4.4	Fine-grained word frequency analysis (Kremmel 2016) (RQ3).....	94
4.5	<i>Discussion</i>	95
4.5.1	Changes in the number of words produced	96
4.5.2	Percentage v. raw scoring systems	97
4.5.3	Adopting a fine-grained approach (Kremmel 2016)	99
4.5.4	Lex30: Possible future changes.....	100
4.6	<i>Conclusion</i>	100
Chapter 5: Collocational changes during SA.....		102
5.1	<i>Importance of collocations</i>	103
5.1.1	The function of collocations	104
5.1.2	Defining collocations	105
5.1.3	Measuring collocations	107
5.1.4	The use of dictionaries to identify collocations.....	108
5.2	<i>Collocation and study abroad</i>	109
5.2.1	Acquisition of collocation knowledge.....	109
5.2.2	Influence of culture and first language.....	111
5.2.3	Towards a new experiment.....	112
5.3	<i>Research questions</i>	114
5.4	<i>Experiment: Collocation and study abroad</i>	114
5.4.1	Method of analysis	115
5.4.2	Results: Overall Performance (RQ1).....	116

5.4.3	Variation by proficiency group (RQ2)	118
5.5	<i>Discussion</i>	120
5.6	<i>Conclusion</i>	121
Chapter 6:	Orthographic changes during SA	123
6.1	<i>Orthography and depth of knowledge</i>	123
6.2	<i>Partial word knowledge and Zareva (2012)</i>	125
6.2.1	Measuring orthographic changes.....	127
6.2.2	Fitzpatrick (2012) and tracking orthographic changes	128
6.2.3	Cook (1997) and the ‘Dual-route’ model	130
6.2.4	Japanese L1 speakers and English orthography	131
6.3	<i>Towards an orthographic experiment</i>	136
6.4	<i>Research questions</i>	136
6.5	<i>Methodology</i>	137
6.5.1	Subjects	137
6.5.2	Task administration.....	137
6.5.3	Analysis and word set classification.....	137
6.5.4	Proficiency groups and partial knowledge	139
6.6	<i>Results</i>	139
6.6.1	Changes in learners’ spelling during SA (RQ1).....	139
6.6.2	The impact of proficiency level on spelling (RQ2).....	144
6.6.3	Patterns of spelling errors and the impact of L1 (RQ3).....	144
6.7	<i>Further discussion</i>	146
6.8	<i>Conclusion</i>	148
Chapter 7:	Wmatrix and changes during SA	150
7.1	<i>Using Wmatrix</i>	151
7.2	<i>Interpreting Wmatrix’s output</i>	153
7.3	<i>Wmatrix, corpus analysis and EFL</i>	156
7.3.1	Analyzing grammatical categories: Lin (2014).....	157
7.3.2	Analyzing key words and semantic domains: Lin (2017)	160
7.3.3	Comparing Lin (2014; 2017) with my present study.....	163
7.4	<i>Research questions</i>	164
7.5	<i>Methodology</i>	165
7.6	<i>Results</i>	167
7.6.1	Comparison of SA data with existing corpora (RQ1)	167
7.6.2	Comparison of SA data with Lin’s 2014 and 2017 corpora	170
7.6.3	Investigating changes within SA data (RQ2).....	173
7.7	<i>Discussion</i>	176
7.8	<i>Conclusion</i>	179
Chapter 8:	Discussion	181
8.1	<i>Productive vocabulary knowledge and Lex30</i>	183
8.1.1	Changes in the number of words produced	183
8.1.2	Fine-grained analysis of word frequency (RQ3)	187
8.2	<i>The effects of SA on collocational behaviour</i>	188
8.2.1	Increases in productive collocational knowledge.....	188
8.2.2	Creating new ways to measure collocations.....	190

8.3	<i>The effects of SA on orthography</i>	192
8.3.1	The importance of spelling.....	193
8.3.2	What orthographic changes were revealed?.....	193
8.3.3	How can an SA experience improve spelling skills?.....	195
8.4	<i>The effects of SA on the acquisition of vocabulary within semantic domains</i>	196
8.4.1	Comparisons of an SA corpus with other corpora.....	197
8.4.2	Comparisons of semantic domain representation.....	198
8.4.3	Future applications of Wmatrix to SA research.....	199
8.5	<i>Adapting Lex30 for future research</i>	202
8.5.1	Variation in the number of responses.....	203
8.5.2	Word frequency lists.....	205
8.5.3	Selection of cue words using norms databases.....	207
8.5.5	Japanese Word Association Database of English.....	208
8.5.6	Lex30: Methods of delivery.....	209
8.6	<i>Discussion summary and conclusion</i>	210
Chapter 9	Conclusion	212
9.1	<i>Some practical implications of this study</i>	212
9.2	<i>Strengths and weaknesses of the research</i>	214
9.3	<i>Semantic Diversity (SemD) and SA</i>	216
9.4	<i>Out-of-class contact: New opportunities for learning</i>	220
9.5	<i>The emotional impact of SA</i>	221
9.6	<i>Final thoughts and conclusion</i>	224
	<i>References</i>	229
	<i>Appendices</i>	253
	Appendix 1: Replication study Lex30 Scoring protocols.....	253
	Appendix 2: Example of Lex30 marking protocol.....	255
	Appendix 3: SPSS Lex30 Analysis : Total number of words.....	256
	Appendix 4: SPSS Lex30 Analysis : Lex30 score (1K+ words).....	257
	257
	Appendix 5: SPSS Collocation Analysis : Overall Changes.....	258
	Appendix 6: SPSS Collocation Analysis : Higher Proficiency Group.....	259
	Appendix 7: SPSS Collocation Analysis : Lower Proficiency Group.....	260
	Appendix 8: Spelling decline examples.....	261
	Appendix 9: Spelling decline then improvement examples.....	263
	Appendix 10: Spelling improvement then decline examples.....	265
	Appendix 11 : USAS Semantic Tagset.....	266
	Appendix 12 : UCREL CLAWS7 Tagset.....	267
	Appendix 13 : Differences at part-of-speech level between the Taiwanese and British discourse in BATTICC (after Lin 2014, p. 311).....	270
	Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison.....	271

List of tables

Table 3.1: Fitzpatrick and Clenton 2010: Longitudinal study score data.....	63
Table 3.2: Fitzpatrick (2003): Longitudinal study score data.....	66
Table 3.3: Replication Study (2016): Study group participants.....	67
Table 3.4: Replication Study (2016): Longitudinal study score raw data.....	69
Table 3.5: Replication Study (2016): Lex percentage scores before and after SA.....	71
Table 4.1: Test administration and number of days between tests	82
Table 4.2: Total words produced by participants during their SA experience	87
Table 4.3: Descriptive analysis: total words	91
Table 4.4: Descriptive analysis: infrequent words	92
Table 5.1: Descriptive Statistics: overall performance	116
Table 5.2: Descriptive Statistics: Higher proficiency group	118
Table 5.3: Descriptive Statistics: Lower proficiency group	119
Table 6.1: Categories of commonly misspelled words (after Tuladhar and Akatsuka 2017, p.99)	134
Table 6.2: Number of responses and misspelled words given at each test time	140
Table 6.3: Percentage of items misspelled at each test time	140
Table 6.4: Groups where identical words appeared (Groups 1 to 4)	140
Table 6.5: Groups with both correctly spelled and misspelled words (Groups 5 to 8)	141
Table 6.6: Spelling improvement examples	141
Table 6.7: Percentage of misspelled words out of total words: High/Low proficiency	144
Table 7.1: Wmatrix Major Categories. (reproduced from Archer et al. 2002, p.2)	151
Table 7.2: SA corpora and time points	166
Table 7.3: Keyword lists: SA corpus versus BNC Samples Informal Written corpus.....	168
Table 7.4: POS: Complete SA corpus versus BNC Samples Informal Written corpus	169
Table 7.5: POS: Complete SA corpus versus BNC Samples Informal Written corpus	169
Table 7.6: POS Comparisons: Lin (2014) , BNC and SA corpora.....	170
Table 7.7: Word frequency and semantic domain comparisons: Lin (2017) , BNC..... and SA corpora	171
Table 7.8: Word frequency keyness: SA (Time 2) and SA (Time 3)	172
Table 7.9: POS frequency keyness: SA (Time 2) and SA (Time 3).....	174

Table 7.10: Semantic frequency keyness: SA (Time 2) and SA (Time 3).....	176
Table 7.11: Selection of Lex30 cues and frequently produced associate words.....	178
Table 8.1: Comparison of the original Lex30 task with an adapted version..... (after Brown 2018, p.101)	192
Table 9.1: Glasgow Norms applied to selected Lex30 cue words.....	223

List of figures

Figure 2.1: Number of international students in 2001 vs 2017.....	12
Figure 2.2: International students studying in the US: 1949 to 2021	13
Figure 2.3: Number of Japanese participating in Study Abroad	14
Figure 2.4: Number of Japanese university students studying abroad	16
Figure 2.5: Meara’s 2005 Model bilingual network	38
Figure 2.6: Vocabulary Test Capture Model (From Fitzpatrick and Clenton 2017, p.19)	47
Figure 2.7 Lex30 sample test with data (from Fitzpatrick and Clenton, 2010, p. 554)	50
Figure 3.1: Relationship between pre- and post-SA Lex30 tests: Replication v	69
Fitzpatrick and Clenton (2010, p.544)	
Figure 3.2: Relationship between pre- and post-SA Lex30 percentage Scores	71
Replication v Fitzpatrick (2003, p.188)	
Figure 4.1: Total words produced at each time point by all 38 SA participants	88
Figure 4.2: Total infrequent words (1K+) produced at each time point by all 38 SA.....	88
Participants	
Figure 4.3: Lex30 task: total words produced over time	92
Figure 4.4: Lex30 task: infrequent (1K+) words produced over time	93
Figure 4.5: Comparison of fine-grained word bands with Lex30 word band divisions	95
Figure 5.1: Mean number of collocations produced per participant	117
Figure 9.1: Number of responses to cue words: 1K and 1K+ frequency levels.....	202

Abbreviations

AH	At Home
BATTICC	British and Taiwanese Teenage Intercultural Communication Corpus
BFP	Brainstorm Frequency Profile
BNC	British National Corpus
CAF	Complexity, Accuracy and Fluency
CANCODE	Cambridge And Nottingham Corpus Of Discourse in English
CANELC	Cambridge and Nottingham E language Corpus.
CCB	Climate Change Bill
CEFR	Common European Framework of Reference for languages
COVID-19	Coronavirus Disease 2019
DIALANG	European development project: DIAgnostic LANguage tests
EAT	Edinburgh Associative Thesaurus
EFL	English as a Foreign Language
ERASMUS	EuROpean community Action Scheme for Mobility of University Students
ETS	Educational Testing Service
EVST	Eurocentre Vocabulary Size Test
GPA	Grade Point Average
G-STEP	Georgia State Test of English Proficiency
HSBC	Hong Kong and Shanghai Banking Corporation
IIE	Institute of International Education
IM	Immersion
JACET	Japan Association of College English Teachers
JANU	Japan Association of National Universities
JAOS	Japan Association of Overseas Study
JASSO	Japan Student Service Organization
JPSS	Japan Study Support
LFP	Lexical Frequency Profile
LoS	Length of Stay
L1	Language 1 (first or native language)
L2	Language 2 (second language)
MEXT	Ministry of Education, Culture, Sports, Science and Technology
NS	Native Speaker
NNS	Non Native Speaker
NIPSSR	National Institute of Population and Social Security Research
OECD	Organization for Economic Cooperation and Development
POS	Part Of Speech
PVLT	Productive Vocabulary Level Test
SA	Study Abroad
SD	Semantic Dementia

SemD	Semantic Diversity
SDM	Semantic Distinctiveness Model
SNS	Social Network System
TOEIC	Test Of English for International Communication
TOEFL	Test Of English as a Foreign Language
VKS	Vocabulary Knowledge Scale
VKT	Vocabulary Knowledge Test
VTCM	Vocabulary Test Capture Model
WAT	Word Association Task

Chapter 1: Introduction

This chapter will present an overview of some of the changes that have affected Study Abroad (SA) programmes and outline some of difficulties experienced in identifying changes in language proficiency. It will introduce a number of measurement tools and briefly describe the selection process of the one most capable of tracing changes in productive vocabulary knowledge before and after an SA experience. In addition to word frequency changes, three further methods of language analysis will also be considered. Finally, the chapter will outline the aims of this thesis and mention some of the challenges caused by the impact of the worldwide COVID-19 pandemic.

1.1 Evaluating Study Abroad programmes

For many years SA programmes have encouraged participants to attend educational institutions in another country and immerse themselves in a different culture. The SA tradition goes back several hundred years to the Grand Tour which was considered to be the high point of any aristocratic education and designed to expand the horizons of the young members of elite British families by introducing them to a wide variety of languages, culture and art (Sanz & Morales-Front, 2018). SA programmes have certainly come a long way since then. The number of students participating has enormously increased and such programmes now come in many shapes and sizes. Some are geared solely around studying and attending a foreign school or university, while others emphasize internships or volunteer experiences. Programmes also vary in how the student is supported, with some having a “host family” situation, while others provide simpler accommodation options. Many have typically allowed participants to spend periods ranging from a single term or semester studying with some extending up to a year or more.

Changes in SA programmes, particularly in Japan, have seen them becoming much shorter in duration. According to McCrostie (2017), writing for the *Japan Times* newspaper, the percentage of Japanese students studying abroad for less than a month increased from 46 percent of the total number of SA trips taken to 61 percent between 2009 and 2015. At the same time the Japan Student Service Organisation (JASSO), which organizes and funds many SA programmes, found in its 2015 survey that fewer than 2,000 Japanese participants studied overseas for more than a

year. In many cases the primary objective of SA is to improve L2 proficiency so the shortening of SA programmes can represent a particular challenge when it comes to their evaluation. Generalized testing using established English language proficiency tests seem incapable of detecting significant levels of change over short periods.

My own university, Nakamura Gakuen, has experience of running SA programmes in Canada and the UK over a number of years. In common with many such institutions in Japan, the programmes it offers have become shorter leading to concern among stakeholders such as administrators, potential employers, parents and students themselves about the benefits they can offer and how to evaluate them. For a long time the university has used the Test of English for International Communication (TOEIC) to evaluate changes in SA participants' language proficiency. Over longer periods there was evidence that the test could track differences in proficiency but changes have become more difficult to determine as length of SA has been reduced (Powers & Powers, 2015).

The broad challenge motivating the studies in this thesis, is to identify a test that might detect changes in learner proficiency over shorter SA periods. As an alternative to using a generalized test to measure learners' proficiency I considered concentrating on a single aspect of language proficiency to see if this offered any chance of success. Researchers, including Laufer and Nation (1999), Maximo (2000), Read (2000), Gu (2003), Nation (2001) and others, have noted that the acquisition of vocabulary is essential for successful second language use. Schmitt (2000) further underlines the importance of vocabulary acquisition emphasizing that, "Lexical knowledge is central to communicative competence and to the acquisition of a second language" (p. 55). Given the important role that vocabulary knowledge may play in overall second language proficiency it makes sense to look more carefully at this aspect.

1.2 Measuring vocabulary knowledge

Many researchers (e.g., Nation, 2001; Schmitt, 2010) conceptualize vocabulary knowledge as having two basic forms which can be evaluated in different ways. Receptive knowledge is typically measured according to the number of words that learners can understand through reading or listening while productive vocabulary knowledge can be taken as the number of words which learners are capable of producing through writing or speaking tasks. Given that the focus of this

thesis will be on tracing changes in vocabulary knowledge during a short-term SA programme the second form of knowledge seems to offer the most promise as research shows that changes in productive vocabulary knowledge are likely to be detectable over the short term (e.g., Fitzpatrick, 2003; Meara, 2005; Fitzpatrick & Clenton, 2010).

As explained in detail in the following chapter, there are a number of measurement tools available for assessing productive vocabulary knowledge. After conducting a critical evaluation of these candidate tests, Lex30, created and developed by Meara and Fitzpatrick (2000), seems to offer the most promise and was chosen for this thesis. There are number of advantages that Lex30 has over other methods which seek to measure the same construct and these will be explained and discussed in chapter 2. The process through which the reliability and validity of Lex30 have been assessed and some examples of some different ways it has been implemented will be described (Baba, 2002; Fitzpatrick & Meara, 2004; Fitzpatrick & Clenton, 2010; 2017).

1.3 Aspects of vocabulary knowledge

Measuring changes in productive vocabulary knowledge entails two key challenges. First, informative samples of learners' productive vocabulary must be elicited before and after an intervention in an efficient way. Secondly, those samples must be scrutinised for evidence of changes in underlying knowledge. A useful approach here is to see vocabulary knowledge as componential, consisting of a number of different aspects. This is in fact a well-documented approach in vocabulary research: many studies are framed around Nation's taxonomy of vocabulary knowledge (Nation, 2001), and several conceptualise depth of knowledge as the acquisition of aspects of knowledge (e.g., Qian, 2002; Schmitt, 2010; Milton, 2013). In the studies presented here, the first challenge is addressed by sampling vocabulary from a number of participants before, during and after an SA experience. The tool used to elicit vocabulary (Lex30) uses carefully chosen cue words, and participants produce associate words for each one. This is repeated at multiple timepoints, thereby creating a set of comparable vocabulary samples for each participant. The words gathered by such a method are then available for further analysis. The second challenge is then addressed, through a number of different methodologies, which are implemented to investigate various aspects of vocabulary

change. Initially this will involve recognizing if there are changes in the frequency of words that SA participants tend to use over time; this was the original purpose of the Lex30 tool and is based on the assumption that learners with larger vocabularies will produce a higher proportion of infrequent words. The samples lend themselves to scrutiny for information about other changes in knowledge too, and other analyses will investigate whether there are changes in participants' collocational knowledge and in their ability to spell individual words. Still further, the data can be used to investigate if participants tend to acquire words belonging to particular semantic groups during their SA experience, or whether there might be a tendency for them to use the same word in increasingly different ways. These changes in different aspects of vocabulary knowledge are outlined below.

Firstly, as SA learners spend longer periods in an L2 environment there is some evidence to suggest an increase in their knowledge of collocations. Collocations (e.g., black coffee, weak tea, terrorist attack, healthy lifestyle) are an important type of formulaic sequence. Formulaic sequences are fixed combinations of words that have a range of functions and uses in speech production and communication (Wray, 2002). Lex30 can attract responses which also happen to be collocates of the original cue words and these might indicate incremental changes in learners' knowledge of collocation during the course of an SA experience. Past studies with Saudi EFL learners studying in the UK show an increase in collocational knowledge usually in proportion to the length of the SA experience (Gobert, 2007; Alqarni, 2017). Analysing vocabulary changes in Japanese EFL learners in a similar UK environment will enable us to identify any corresponding improvements in collocation acquisition and use.

Secondly, we can investigate how SA learners' spelling (orthographic) accuracy varies over multiple test times. Past uses of Lex30 show that cues will sometimes elicit the same response at consecutive test times which might be useful for analysis (Fitzpatrick, 2012). It may be possible to see how well individual words are known and whether there is a particular pattern for improvement (or deterioration) over time. A further interesting point is that patterns of such orthographic change can give an insight into writing issues connected with particular L1 groups. For instance, Japanese EFL learners will tend to make different spelling errors to Spanish EFL learners. Further research might reveal further details of what

examples of partially familiar vocabulary, also termed *frontier words*, are present (Zareva, 2014).

Thirdly, an SA experience gives learners the opportunity to encounter concepts, and therefore vocabulary items, in previously unfamiliar domains. Scrutiny of the kinds of newly acquired items produced after an SA period can provide insights into whether certain areas of vocabulary growth are encouraged by SA programmes. Examples could include descriptions of places, education, people and travel. It is challenging to elicit such information from a relatively small vocabulary sample, but the application of semantic analysis tools such as Wmatrix (Rayson, 2003) may detect changes in the semantic mastery of certain domains. Such information might prove useful in the creation and development of materials for new SA programmes.

Finally, it might be of use to evaluate some future directions that research in vocabulary changes may take us. For instance, one aspect of vocabulary growth which can be particularly difficult to detect is the developing knowledge of the context in which known items can be used. Adelman et al., (2006) carried out research which showed that word frequency may not be the only organizing principle underlying lexical access. They found that counting the contexts in which certain words appear may give a better quantitative fit to human lexical decision than merely counting their raw occurrences. To put it in a different way, it is not just the number of times an individual encounters a word that is important, but also the quality and diversity of those encounters. Individuals who read more widely will experience more of these opportunities, which may help to explain why they are better readers. Hsiao and Nation (2018) conducted experiments with children which involved reading and vocabulary knowledge and found that high diversity words were responded to faster and read more accurately than low diversity words. Their findings demonstrated that contextual variability contributes to word learning and the development of lexical quality, independent of the frequency of occurrence of individual words. Analysis of vocabulary samples from SA participants has the potential to enable us to discover whether the increase in learning contexts that SA is more likely to offer can help improve the process of vocabulary acquisition of certain words.

We also might consider the effect of both the psychological and environmental impact that SA may have on language acquisition. Some previous

studies have found that students can experience psychological changes which are demonstrated by lower levels of mental function while studying abroad (Hunley, 2010; Savicki, 2013). This psychological impact can manifest itself in the forms of vocabulary acquired by learners and how these might change during the course of an SA programme. In addition, analysis of different environments and of the variations in opportunities available for out-of-class language contact may have an important effect on vocabulary gain (Briggs, 2015). Thinking about such issues may help educators to enhance students' experiences while abroad and support the provision of intercultural training and more supportive SA environments.

1.4 The organization of this thesis

This thesis will identify a variety of different changes in vocabulary knowledge among Japanese EFL learners participating in a short-term SA programme in the UK. The literature review in chapter 2 will begin by looking at the history and trends of SA programme development both internationally and in Japan. It will attempt to explain the recent tendency for reduced length SA programmes and will underline the need for more sensitive measuring instruments to evaluate them. One example of a productive vocabulary task, Lex30, has the potential to meet this need so the literature review will review and discuss a number of validation studies connected with this measurement tool. Several examples of its use in studies with EFL learners will be covered as well as alternative methods of scoring and analysis.

Chapter 3 reports the replication of two experiments (Fitzpatrick, 2003 and Fitzpatrick & Clenton, 2010) where Lex30 has been used to measure changes in vocabulary frequency during SA programmes. Chapter 4 of the thesis presents a more substantial longitudinal experiment where Lex30 data has been collected at a number of time points before and after an SA programme. Chapters 5 to 7 describe alternative analyses starting with a look at how the number of words can change according to word frequency bands, before going on to examine changes in collocational use, orthography and semantic grouping. Each of these chapters (5 to 7) will start with a short literature review providing a more detailed and substantial background to each particular area of research than covered in the general literature review (chapter 2).

Chapter 8 will draw together findings from the experiments reported in the thesis, to identify key changes in vocabulary knowledge resulting from an SA

experience. The chapter will finish with a summary of the results of each stage or chapter of the thesis. It will describe the contributions that each step can make towards the field of vocabulary knowledge. Finally, chapter 9 will propose three potential areas for future research including: (1) how the number of different learning contexts may affect how new words are acquired, (2) how differences in the SA environment and out-of-class language contact opportunities may affect learning outcomes and (3) how methods of psycholinguistic analysis might inform us how the emotional impact experienced by SA participants may affect their ability to acquire new language.

1.5 The impact of COVID-19 on study abroad research

The effects of the worldwide pandemic COVID-19 on SA have been profound. Starting in February 2020, a large number of SA programmes were cancelled worldwide as governments fought to lower infection rates by limiting the movement of people. The repercussions have been particularly serious in Japan. According to a Japan Association of Overseas Study (JAOS) 2021 survey on its membership organizations in 2020, the number of students going abroad was 18,374, a 76% decrease from a year earlier. For my own research and with the studies presented in this thesis, COVID-19 has presented its own challenges.

In the year before the COVID outbreak, I carried out a longitudinal experiment which sampled productive vocabulary knowledge at various timepoints before and after a short-term SA programme. I had intended to collect further data in subsequent experiments, using a larger number of participants, stratified by proficiency, which could then be used to make comparisons between Lex30 and other similar tests measuring the same construct. These tests included the Lexical Frequency Profile (LFP; Laufer & Nation, 1995); the Productive Vocabulary Level Test (PVL; Laufer & Nation, 1999) and the Vocabulary Knowledge Test (VKT; Paribakht & Wesche, 1993). This research would also have built on Walters (2012) who explored the differences between SA participants' ability to simply recall individual words or whether they could use vocabulary meaningfully and appropriately.

The subsequent impact of COVID-19 on this planned data collection created some new challenges. Since it became difficult to carry out any further data collection, I realized there was a need to develop new methodologies capable of

mining the data I had already gathered in a number of new and creative ways. By the beginning of 2020, when the pandemic was starting, I had already conducted an experiment which replicated two earlier studies (Fitzpatrick, 2003; Fitzpatrick & Clenton, 2010) and had found evidence to suggest that that Lex30 was capable of detecting longitudinal changes in productive vocabulary knowledge. I had also collected data for a more extensive study where the same task had been administered on four occasions to a larger sample group. When it became evident that COVID would restrict capacity for further data collection, I changed my study design to focus on the data already collected.

Chapter 2: Literature review

This chapter starts out by considering the general history and growth of Study Abroad (SA) within the context of language education in Japan. It covers many of the recent study trends in that country particularly the move towards short-term SA programmes. In addition it will look at the evaluation of the effectiveness of SA and examine which particular aspects of language might undergo the most change. Identifying vocabulary knowledge as one area showing considerable promise, the chapter will go on to assess a number of measurement tools with which to carry out data collection. One particular tool, the Lex30 productive vocabulary test, emerges as the most appropriate for this task. The way this instrument works is explained and past validation and reliability studies of its performance are reviewed.

2.1 Study Abroad (SA) Definition and history

Study Abroad (SA) can be described as an academic activity that expands and develops world views while helping participants acquire a foreign language and cultural knowledge. Any sort of travel has educational potential, whatever its inspiration and purpose, and what is learned in large depends on how individuals interact with the new world around them. Students travelling abroad typically have the opportunity to observe, participate and communicate in classrooms, workplaces and homes and experience a wide range of personal relationships, service learning, or commercial interaction (Kinginger, 2009).

SA programmes come in many forms with different objectives, academic foci and stakeholder expectations. A typical university programme, for example, might allow a student to spend a single term or semester studying abroad while other programmes could include a year or more of overseas study. Some are geared solely around studying and attending a foreign university, while others emphasize internships or volunteer experiences. The SA experience can also vary in how the student is accommodated. Firstly, there is the notion of reciprocal SA involving participants from each country who live at the other's respective houses or secondly, where the participant either simply stays with a "host family" or makes arrangements to stay at a university dormitory or private apartment. There are also SA programmes for high school students and college graduates. High school students, due to their age, are often required to live with a host family or in a more highly supervised

living situation. With university graduates a recent trend has been the opportunity to teach abroad, frequently as part of a master's degree programme to become a teacher, usually of a foreign language, in their home country. Other graduates participate in ongoing research at foreign universities while pursuing a graduate education. On a less formal level SA is sometimes undertaken by individuals of almost any age who are not involved in full time education but are keen to participate for their own interest or perhaps as a retirement project.

Stakeholders such as governments, academic institutions, parents and students themselves, are likely to consider whether a particular SA experience will justify the investment of time and financial resources. Programme cost and length together with corresponding predictions about language proficiency improvement and how far relevant skills or work experience are gained are likely to be key decision-making factors.

On a global scale there are a number of different approaches that countries take concerning issues with SA. A 2004 Organization for Economic Cooperation and Development (OECD) report: *Internationalization and Trade in Higher Education* looks at the growth and nature of SA describing four kinds of approach or national policy. These include (1) mutual understanding, (2) revenue generation, (3) skilled migration and (4) capacity building. The first of these, the mutual understanding approach, describes providing encouragement and stimulus to academic research through exchange programmes such as the Fulbright Scholarship in the United States and ERASMUS¹ in the European Union (EU). The second, the revenue generation approach, refers to the way in which some countries, for example the United Kingdom, Australia or New Zealand, promote the services of their higher education system to fee-paying students from foreign countries in order to develop their own education infrastructure in an attempt to gain a larger world market share. The third, the skilled migration approach, aims to attract highly skilled students to first study and then to remain in the host country after their course finishes both to counter the negative economic effects of an aging society and also in an attempt to stimulate academic research. Examples include Germany and more recently Japan. The final, capacity building approach, refers to how some countries, such as Sri Lanka or Cyprus, encourage students to first study abroad to gain useful and important skills and then return to help build and improve society and the education system in their

¹ ERASMUS: EuRopean community Action Scheme for Mobility of University Students.

own countries. This is particularly important for countries where the demand for higher education may outstrip the supply.

The four approaches described are not mutually exclusive and may overlap and operate in many combinations. Many of the SA programmes described later in this thesis follow the third approach and involve young Japanese learners studying in the UK. Instead of remaining for an extended period after completing their studies, however, they typically return to their home country to use their newly gained linguistic skills to make an important contribution to academic research and also to further help strengthen trade and economic ties between Japan and the outside world.

2.1.1 The history and growth of SA

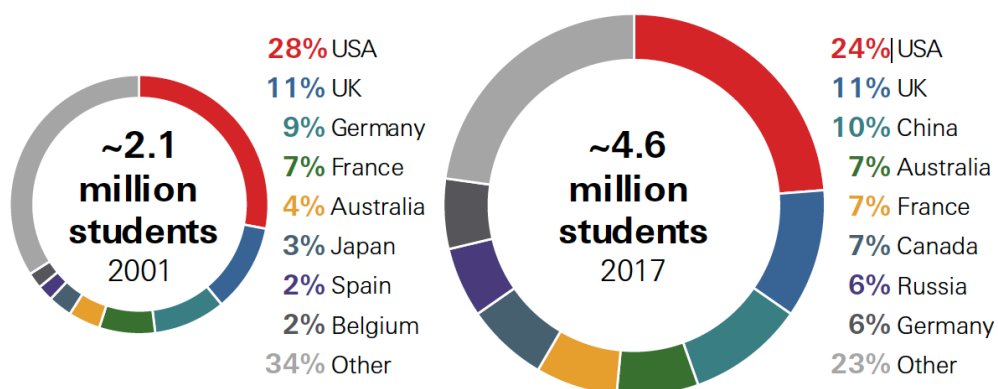
SA is by no means a new phenomenon. Education exchange of one form or another has carried on across national borders for thousands of years. Nor has it been only confined to language skill improvement. Centres of civilization and learning, firstly in Mesopotamia, then in Greece and Rome attracted migrants across the ancient world who were keen to improve their minds and prospects across a wide range of disciplines. The middle ages saw talented students across Europe gravitating towards ancient seats of learning at universities like Bologna, Oxford and Salamanca to study subjects like medicine, theology or law. Until more recent times, however, study abroad was reserved for very privileged and the royal elite. Then there came the first signs of expansion in the eighteenth century when figures like Emmerich de Vattel (1844), a Swiss diplomat, argued for the exchange of professors among various nations as many thought that this would promote the peace and prosperity of all. Another important figure was Marc-Antoine Jullien (1792), a French educator, who saw the potential benefits of sharing ideas and developing mutual trust among academic institutions (Wolhuter, 2016). The peace congresses following the Napoleonic wars further helped to create the groundwork for the field of international education that we are familiar with today.

The aftermath of the world wars in the twentieth century saw an acceleration of efforts to promote students' awareness of the world outside their own national borders. This is shown by, for example, the United States' government's establishment of the Fulbright programme in 1946 which sought to humanize international relations by creating better communications and trust between nations. Since its inception the programme has supported more than 200,000 students who have visited 150 countries.

Until the onset of the COVID-19 pandemic, international higher education or SA continued to thrive. Smith (2017) used OECD statistics to show that new attitudes had developed. A survey looked at 8,000 parents with sufficient financial resources in 15 countries and showed that 42% of them considered sending their child to study abroad, compared to 35% of parents a year earlier. Parents in Asian countries were among the most “outward looking” reveals the report, with the three top countries in the world where parents were considering a university education abroad for their child, being India (62%), Indonesia (61%), and China (59%). The report also highlights the fact that Asian students accounted for 53% of all students studying abroad worldwide. Figure 2.1 (HSBC, 2017) shows the growth of international students across a range of countries. A total of 4.6 million higher education students studied abroad in 2017, a significant increase from 2.1 million in 2001. As for destination countries the United States attracted around 1,079,000 international higher education students, the United Kingdom around 501,000, China 443,000, Australia 328,000 and France 324,000. When looking at Figure 2.1 two points are worth noting. Firstly, there is a large increase in the number of Chinese students studying abroad from less than 1% of the total number of international students in 2001 to more than 10% in 2017. This is perhaps a reflection on the rise in the economic development of that country. Secondly, in the case of Japan, it would seem that the other countries shown have attracted a greater number of international students in proportion to the total of 4.6 million which might account for the lack of data for the country for 2017 (it does not appear on the second list in Figure 2.1).

Figure 2.1

Number of international students in 2001 vs 2017



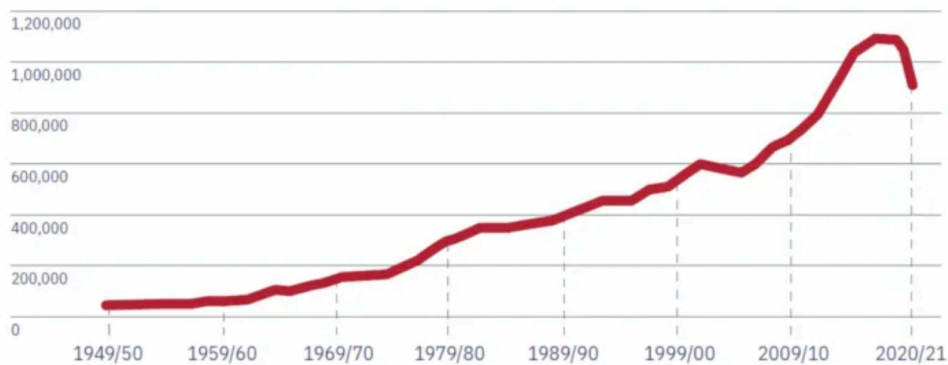
From HSBC Global factsheet (2017): Globalisation of higher education Source: OECD

2.1.2 The impact of the COVID-19 global pandemic

From 2020 SA has been transformed due to the onset of the global COVID-19 pandemic. The impact has been profound with significant restrictions still remaining in place in some countries. The New York-based Institute of International Education (IIE) reports a survey of 860 U.S. institutions of higher education which is shown in Figure 2.2 (Yifan, 2021). It shows that for the 2020-21 academic year, the number of international students, mostly from Asia, fell 15% to 914,095. Further information, which is not given on Figure 2.2, shows that the number of Japanese students undertaking SA in the United States suffered the biggest percentage drop among the top 20 countries in 2020-21, with a 32.9% decline. According to the IIE this was partially due to the fact that the majority of students from Japan were engaged on short-term programmes which seemed to be more widely affected than longer degree-level ones.

Figure 2.2

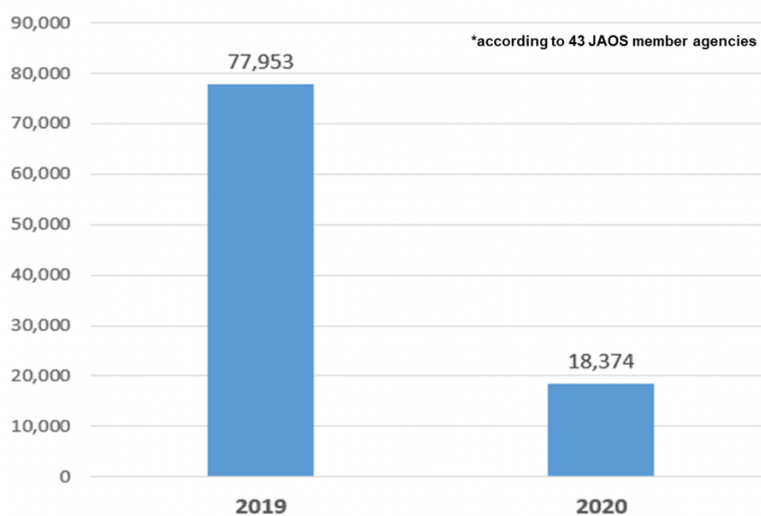
International students studying in the US: 1949 to 2021 (Source: Yifan Yu, Nikkei Asia 2021)



Another survey by a study-abroad industry association based in Japan, the Japan Association of Overseas Studies (JAOS), has further revealed the extent of the impact the worldwide pandemic has had on Japanese students (JAOS, 2021). As seen in Figure 2.3 the number of Japanese students pursuing study-abroad opportunities in all countries saw an unprecedented 76% drop attributed to COVID-19 between 2019 and 2020. Although by the second half of 2021 there were some early signs of recovery, many universities including a number in Japan, had cancelled or postponed their SA programmes.

Figure 2.3

Number of Japanese participating in Study Abroad (From JAOS report 2021 p.1)



In response to the pandemic there has been some growth of online exchange programmes and this has been seen by many as a useful alternative. Some of these programmes have proven effective, and they are still capable of providing students with a valuable experience. Benefits include much reduced costs which makes it easier for more students to participate and also enables them to continue with their regular life activities in their home countries. It is still too early to predict how SA might change in the post-pandemic era. While there is some probability of SA returning to pre-pandemic levels it also seems likely that online programmes will take on an increasingly important role.

2.2 Japan and Study Abroad (SA)

In Japan the practice of studying English has had an ambivalent relationship with SA. There has always been a distinction between English for practical purposes and English for academic uses which is particularly relevant with regard to university entrance examinations (Butler & Iino, 2005). With the opening of Japan to the West and the start of the Meiji era in 1868, rapid modernization of industry and society encouraged the practical use of English as well as other European languages to gain access to new ideas and innovation. English became the main foreign language taught in schools at the end of the nineteenth century. With the appearance of succeeding periods of nationalism however, especially during the 1904-5 Japan-Russo and Second World Wars, the role of practical English was diminished

although it still maintained its role as academic gatekeeper for access to higher education. With the reorganization of the education system following the Second World War and occupation by the United States, English regained some importance as a means of practical communication and in serving to support industrialization and modernization (Dougill, 2008).

Towards the end of the twentieth century English language education in Japan was still the focus of some controversy. Kinginger (2009) described changes of successive governments' policies and views on the role of English. Opinions have varied from those suggesting that it be adopted as the country's second official language to some which support the enforcement of policies which attempt to separate it from the underlying cultures of English speakers, using it only to serve as an instrument for transmitting cultural, economic messages from Japan to the outside world. All these changes have had an important effect on SA programmes and the way in which they are conducted. The Japanese government's Ministry of Education, Culture, Sports, Science and Technology (MEXT) for the time being includes cultural education as a core element in primary and secondary education. However, the government's current policy continues to change but at least seems to be heading towards promoting a level of functional proficiency for most students (MEXT, 2018).

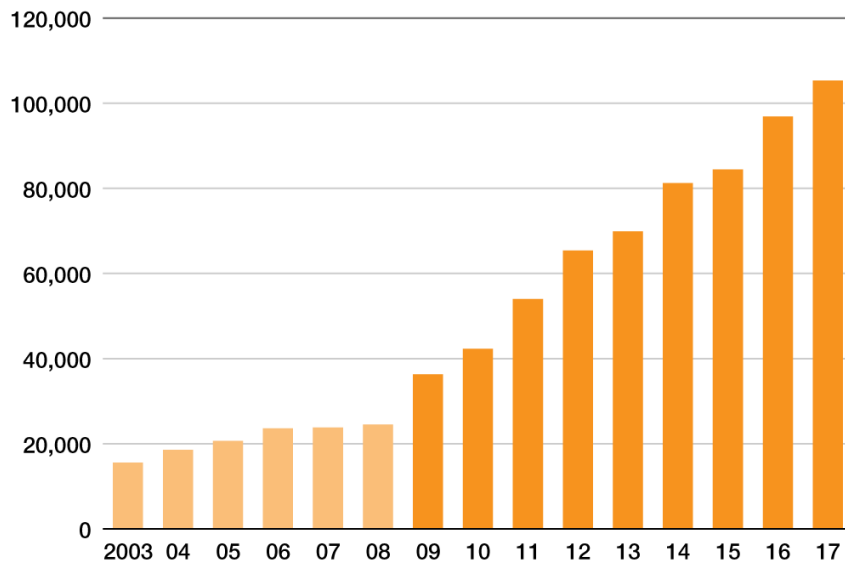
2.2.1 Japanese Study Abroad (SA): Recent trends

Statistics relating to the number of Japanese students studying abroad are not entirely clear cut and seem to differ according to how the data, gathered by different organizations within Japan, is interpreted. McCrostie (2017) explains that some figures show that increasing numbers of Japanese are going abroad to study but that a closer look at the data may reveal that looser definitions of SA may be inflating the numbers. For instance, a greater number of people participating for shorter periods in less formally organized programmes might be responsible for this. MEXT surveys show the number of Japanese enrolled in overseas universities at a peak of 82,945 in 2004 but falling to 60,138 in 2012 and then 53,197 in 2014. This represents a drop of 36 percent (MEXT, 2018). In order to counter this sharp decline MEXT set a goal of doubling study-abroad totals by 2020, raising the number of university students studying overseas from 60,000 to 120,000 and the number of high school students from 30,000 to 60,000. The thinking behind this new policy on SA has come about in large part because Japanese firms, hoping to strengthen their overseas operations, are

struggling to find enough workers with the required language abilities and international experience. To help meet this goal, in 2014 MEXT launched “Tobitate! (Leap for Tomorrow),” an initiative whereby government money and corporate donations would provide funding 10,000 scholarships for university and high school students to study overseas.

Figure 2.4

Number of Japanese university students studying abroad



Nippon.com based on data from the Japan Student Services Organization (JASSO)²

Data from the Japan Student Services Organization (JASSO), a quasi-autonomous agency responsible for non-government scholarships and student loans, on the other hand, shows some increase in the number of SA participants (JASSO 2016). Figure 2.4 shows 84,456 Japanese studying at overseas universities in 2015, up from 36,000 in 2009 and rising to more than 105,000 by 2017. Some discrepancy with the previous MEXT figures perhaps is explained by JASSO counting students participating in short-term exchanges, sometimes as short as eight days, at overseas universities and colleges. Such exchanges might include short-term intensive language courses, cultural exchanges and research trips. Another factor that may be responsible is that JASSO uses originally recorded and scheduled SA information,

² For the period 2003 to 2008 data was only compiled on students who were studying abroad under exchange programmes; whereas from the 2009 academic year the survey also targets students who enrolled at non-partner overseas institutions.

regardless of whether a student returns home early. For example, this means a student who actually intends to spend a year abroad but then returns after four months still gets recorded in the one-year-abroad category. According to McCrostie (2017), SA trips lasting under a month account for nearly all the increase in JASSO's numbers. He quotes JASSO's 2015 figures which state that between 2009 and 2015 the percentage of Japanese studying abroad less than a month increased from 46 percent of the total to 61 percent while less than 2,000 Japanese studied overseas for more than a year.

Finally, figures from the Japan Association of Overseas Studies (JAOS) which is made up of 66 private and public-sector organizations, seem to predict a bright future. Its 2017 survey shows that the number of Japanese studying abroad, including working adults, exceeds 200,000 (JAOS, 2017). This figure includes data which had not been previously included by other surveys on Japanese people studying abroad. JAOS combines the statistics obtained by both MEXT and JASSO along with data gathered from its own member organizations. McCrostie (2017) suggested that if the likely number of students who go abroad via education agencies that are not members of JAOS are added, the original estimate is likely to rise even further. Although JAOS does admit some difficulty in obtaining a final precise figure, the wide range of statistics quoted by some of the different organizations involved in SA illustrate a lack of consistency in data gathering and difficulties in its interpretation.

Shimmi and Ota (2018) reported that there has been a sharp increase in the number of students participating in what they term as "super-short-term" SA programmes lasting from one week up to one month. Quoting JASSO they say that the number of Japanese students who participated in such programmes more than tripled between 2009 and 2016, increasing from 16,873 to 60,145. Government-funded and JASSO scholarships are partly responsible for this rise as they tend to be awarded more often to short-term SA applicants. At the same time the same scholarships provide comparatively little support for students who seek academic degrees at foreign universities particularly at undergraduate level. The number of recipients of JASSO scholarships for short-term SA increased from 627 in 2008 to 22,000 in 2017. The Tobitate "Young Ambassador Programme" is sponsored by both the government and private companies and provides support for students studying

abroad for periods varying from 28 days to two years. By 2017 about 3,000 university students had received SA funding.

Likely future trends of overseas study are that the reforms of university entrance examinations currently taking place in Japan will encourage more students to go abroad at a younger age. Changes include a greater emphasis on communicative and practical skills using tests based on the Test of English for International Communication (TOEIC) and the Test of English as a Foreign Language (TOEFL). There is a growing recognition that many of the necessary knowledge and skills for future examination success perhaps can be taught more effectively in an SA setting. For both undergraduate and graduate university students as well as working adults the continuing globalization of many Japanese companies and increasing numbers of tourists visiting Japan from overseas is also likely to stimulate demand in SA.

2.2.2 Japanese Study Abroad (SA): Future challenges

Although statistics mentioned above seem to describe a positive view of Japanese students travelling to study overseas there are a number of concerns for policy makers. Shimmi (2012) described some of the structural issues affecting numbers of Japanese students. Firstly, she pointed to how the demographic shift in Japanese society has affected the number of potential SA programme participants. The number of 18-year-olds in the Japanese population peaked at 2.49 million in 1966 as the post-Second-World War baby-boom generation reached that age. The numbers increased again in the 1980s, but after reaching a figure of 2.05 million in 1992, when many of the baby boomers' children finished high school, Japan's population of 18-year-olds fell again, decreasing to 1.52 million in 2001 and 1.21 million by 2015. The National Institute of Population and Social Security Research (NIPSSR) predicts that this number will continue to shrink, falling below the one million mark to roughly 990,000 in 2031. Multiple factors for this decline in the population may exist including rising financial burdens for child-rearing which discourage many younger Japanese people from having families and the increase in the number of younger workers who are unable to secure stable permanent jobs and therefore end up with less financial stability.

Secondly, the continued expansion of Japanese higher education and enrolment opportunities has provided additional opportunities and capacity for in-

country language learning and this has perhaps contributed to the decline of Japanese students who feel the need to study abroad. This is particularly important when considering that fact that many local institutions are now beginning to offer classes in English. The Japan Study Support (JPSS), an online information site, lists several hundred courses including those even held at the prestigious Tokyo, Keio and Sophia Universities (JPSS, n.d). In addition to overseas students, the number of Japanese participants has also steadily increased as they are seen by some as more economical alternative to spending a lengthy and usually more costly time overseas. Many people interested in SA can be discouraged by financial concerns especially in view of high tuition and living fees for study. According to Nippon.com, another online Japanese news and information site, in the United States and some other English-speaking countries, for example, total costs of SA can average from ¥2 million (£14,000) to ¥5 million (£35,000) per year and these figures continue to rise (2019). Measured against this is the fact that disposable income in Japan has declined over the last two decades turning study abroad into a significant economic hurdle for many students and their families.

A third factor could be a possible conflict with students' search for jobs. Many Japanese companies prefer to hire new graduates and a number of different factors will decide when firms release recruiting information and begin the selection of candidates. The ideal window for SA is thought by many university students to be during the third or fourth year of typical four-year university course and usually has the aim of increasing language skills or accumulating knowledge in a specialist field. This period, however, occurs at the same time as students begin their search for new jobs and this fact alone may discourage many from taking up overseas study opportunities. Perhaps new government policies affecting recruitment, introduced from the 2018 academic year (April 2018 to March 2019), might encourage some companies to start their drive to seek suitably qualified candidates at a later stage of students' university lives. Even starting from the summer of a student's third year of university study might increase the time available for SA. It is probably too early to say whether such government guidelines will work as many companies post information early or have internship programmes during summer vacation in the third year. It seems likely that the clash between SA and students' employment prospects will remain for some time to come.

A fourth factor are problems caused by the motivation and anxiety of potential SA participants. Many universities in Japan have been infamously described as a “motivational wasteland” (Berwick & Ross, 1989, p. 207) and have student bodies who are criticized as being inward thinking, risk averse, and not interested in the world outside of Japan (Asaoka & Yano, 2009; West, 2015). Ota (2011) reinforced this notion by suggesting that younger generations of Japanese people have an apathetic attitude towards living abroad due to the cultural and linguistic challenges that are expected, alongside a belief that foreign countries are fraught with dangers. This perception of danger and safety may lead to the tendency to avoid uncertainty, resulting in students who are unwilling, unmotivated, and poorly prepared for study abroad.

A final factor is linguistic anxiety. Despite spending years learning English at school and taking numerous language exams, many students lack confidence about getting by in everyday life in a foreign country let alone studying regular classes. One study revealed that a majority of Japanese university students (55.9%) believe that their English education was not useful (Hirai, 2014), with a claim that “the Japanese education system lacks teaching students how to use English in their daily lives and business scenarios” (2014, p. 2). Research also reveals that there can be a weak association between some students’ perception of their own future prospects and English or other L2 language speaking target groups (Miyahara et al., 1997). This is a concern for stakeholders such as potential employers, as this seems to discourage a willingness to integrate with the wider international community (Knight, 2004) despite a desire to acquire the international L2 skills needed to compete in the global job market (Yonezawa, 2010). Limited practical experience of using English is often seen as an important issue by many. Connected with this is that more exacting academic requirements can also be a difficult hurdle for Japanese students who hope to study for an extended period overseas or at a more prestigious institution. In particular, the new system of the Test of English as a Foreign Language (TOEFL) which is Internet-based, includes a speaking section which seems to be a particular challenge for Japanese students. According to the Educational Testing Service (ETS) TOEFL scores for Japanese students in 2017 rated only 27th among 30 Asian countries that were surveyed. With short-term SA programmes there appear to be fewer academic barriers to cross which might, in part, account for their recent popularity.

2.3 Evaluating SA programmes

Clear methods of evaluation and assessment have an important role when making the decision to run effective SA programmes (Parker, 1999). Stakeholders including governments, universities, employers, parents, and students themselves need to be able to assess the benefits of SA programmes so it would seem important to know what their tangible results are likely to be. Improvements in language test scores, GPA (Grade Point Average) or advantages in securing a desirable job might be important considerations for the student while, from a university's point of view, it has to be decided whether the programme is valuable enough for the time and resources used. Stakeholders require some sort of evaluation. Tanaka and Manning (2018) identified four main stakeholders and assessed the value of SA's contribution for each. They are (1) government, (2) universities and institutions, (3) students and their parents and (4) employers.

The government has become an important SA stakeholder particularly in the case of Japan. Currently MEXT includes cultural education as a core element in its primary, secondary and higher education programmes. It offers a variety of financial support including direct sponsorship of "Tobitate" scholarships and provides additional funding to organizations like JASSO and JAOS to order to increase the numbers of high school and university students studying abroad. The government aimed to double the number of students travelling overseas from 60,000 in 2010 to 120,000 by 2020 in the case of university SA programmes and from 30,000 to 60,000 with high school SA (Japan Association of National Universities; JANU 2016). Evaluation is an important part of providing accountability to the tax-paying public. The question that will be asked is whether the benefits of government funded SA measured in terms of the personal development, linguistic improvement and increased employability of participants can justify the resources used.

Universities and other educational institutions are continuing to find that a number of their students wish to travel abroad but that financial restraints have affected many of them. As a result many universities are beginning to offer special assistance to enable more students to participate in addition to funding from other sources like MEXT or JAOS. Universities are keen to do this perhaps because SA programmes, showcased in open campus events, seem to attract much interest from prospective students which in turn can lead to an increase in the number of applications. SA can also lead to international partnerships between institutions

which can be beneficial for both sides, for example in some cases there are exchanges in teaching staff which can both infuse teaching departments with new ideas and increase the diversity of courses available. Reciprocal student exchanges can increase the opportunities for cross cultural interactions for a wider section of the student population. This has been achieved by bringing students from universities in other countries to both learn languages and culture and also make their own contributions to their host communities. Finally, SA has the potential to increase the employment rate for graduating students thereby raising the profile of the home institution. The question that many universities will ponder is whether a higher number of students undergoing SA will result in raising their university's domestic and international profile which will then, in turn, attract more funding and a greater number of applications.

Students themselves find that SA can enhance their capacity for self-expression and interaction with others. It can expand their ability to analyze issues from a much broader perspective and go a long way to build the confidence to establish long-lasting international friendships. Students can increase their appeal and employability by emphasizing their SA experience on their Curriculum Vitae (CV) and during interviews in the process of seeking a job on graduation. Parents, who will likely be providing most of the financial resources for their children's study, have a strong interest in their future happiness and having them succeed in developing their independence and securing stable employment. For students the main question about SA will be whether the overseas experience will be an enjoyable one, will it produce tangible results like an increase in English language proficiency and finally will it increase the likelihood of obtaining a better job.

From the employer's point of view the importance of SA is clear. There are 20,000 Japanese companies with overseas branches and this number continues to increase (Kuno, 2014). More companies are looking for employees with language abilities and overseas experience. Their view is that such applicants have already demonstrated an ability to adjust to an unfamiliar environment, endure certain hardships and successfully communicate with people that are different from them (Benson et al., 2013). An important point that should be made here is that in the past the focus of SA has been on the development of language ability over the long-term, but more recently the shift is tending to move towards shorter programmes. An additional factor is that with many SA programmes, maintaining the sole emphasis

on language development is becoming less common which seems to match the view of many companies. Kinginger (2013) reported that out of 596 companies she surveyed, 86.6% were looking for well-developed communication skills when selecting new employees and not necessarily foreign language proficiency. Her research revealed that many companies consider the fact that even a short time spent abroad can be useful for developing the ability to perform certain speech acts, such as closing and opening conversations and selecting appropriate politeness markers. It is entirely possible that short-term SA can be sufficient in many cases to increase understanding of how language elements, which is sometimes partially already known by learners, can be used in the correct context by providing opportunities for practice in real world situations. Such social and language etiquette can play an important role particularly in service industries. For the employer SA can improve increased interpersonal skills and a broadened understanding of international business practices (Orahood et al., 2004).

A further point is that SA can expose learners to the diversity of English around the globe and engage them in critical discussion about global English (Dewey, 2012). Raising awareness of and increasing positive attitudes toward the increasing number of English varieties can serve to emphasize the need to focus more on communicative strategies. This concept of ‘Global Englishes’ and how the awareness of the diversity of English (Galloway & Rose, 2018) is challenging traditional approaches to ELT is more fully discussed later in this thesis in sections which relate to the design of Lex30 and specific language items produced by SA participants.

For all stakeholders involved evaluation plays an important role in the creation and administration of an effective SA programme (Parker, 1999). They will usually want to see the results of an investment which involves a considerable amount of time and effort. Tangible benefits such as improvements in language test scores and university performance measures, increases in learner motivation and attitude, changes in a university’s profile and widening educational opportunities and even advantages in securing a desirable job might all be considered important considerations. From a linguistic perspective and the research carried out in this thesis the evidence of the effectiveness of studying abroad is interpreted as the measured development of participants’ foreign language ability.

2.3.1 At Home (AH) versus Study Abroad (SA)

Intuitively SA's main advantage is that it provides an environment more conducive to language learning than normal classroom study. For that reason a common approach of many evaluation designs is to compare the language proficiency of SA groups with At Home (AH) groups with the latter group acting as an experimental control. Both learning contexts are so different from each other that there are likely to be differences in the way language is acquired. SA provides many more opportunities for target language exposure and production so it is therefore probable that differences in language development between the two groups are due to the increased quantity and quality of language input and use during the SA experience. A number of studies have highlighted the differences between the two groups and found significant language gains with SA compared to AH (Freed, 1995; Collentine & Freed, 2004; Freed et al., 2004b; Llanes, 2011 and others). Freed's (1995) participants, for example, studied French as an L2 in two learning contexts, AH and SA. She assessed the participants' fluency by means of native speakers of French who were trained to evaluate the participants speech samples by rating the speech samples on a linear scale. The results led Freed to conclude that "students who have lived and studied abroad were found to speak more and at a significantly faster rate (1995, p. 141)." Freed et al.'s (2004b) study went further and compared the oral development of three groups in three different learning contexts: AH, SA and Immersion (IM). The authors looked at oral fluency in depth by examining a number of speech parameters and found, to their surprise, that the IM groups made most gains. This was attributed to the fact that IM students reported writing and speaking the L2 more hours per week than either of the two other groups. With vocabulary acquisition there have been a number of comparative studies between AH and SA groups (Collentine, 2009; Dewey, 2008; Foster, 2009; Llanes & Muñoz, 2009; Llanes, 2010 and others). For example Foster (2009) found that AH participants used single and general words whereas participants in the SA group used more narrowly defined lexical choices. Foster concluded that living in the target language environment resulted in an enriched lexicon that made the participants sound more native like and more natural. Llanes and Muñoz's (2009) study found that even after three or four weeks abroad the SA group made far fewer lexical errors. There seems to be a general agreement among many researchers on the fundamental role of SA and its contribution towards language development: "One of

the most important variables that affects the nature and the extent to which learners acquire an L2 is the context of learning” (Collentine, 2009, p. 218).

Despite the apparent usefulness of AH versus SA experimental designs there is a growing realization that they can introduce a number of confusing learner-level variables in spite of attempts made by some researchers to limit pre-existing differences between study groups. While it is true that some differences between participants of AH and SA groups can be reduced by administering preprogramme measures, many studies make no such allowance. Some have examined whether students who elect to study abroad are different from their AH counterparts. They often found that they are likely to differ in a number of important ways “such as motivation, aptitude or attitude” (Grey, 2018, p. 49). Rees and Klapper (2008) investigated some of the difficulties of comparing at-home groups and SA groups that may have very different backgrounds and motivations and suggest that the problem is likely to be exacerbated with self-selection bias for participants wanting to attend SA. Kim and Lawrence (2021) found that different individuals’ level of motivation also seemed to affect their success in acquiring a new language during SA. These studies seem to show that it is not just the SA *context* that differs from AH but the SA *learner* as well. It may well be the case that present AH versus SA designs cannot completely avoid these potentially confounding factors which makes considering AH a reliable comparison group for AH conceptually unsound (Grey, 2018). Unlike laboratory studies in which random selection of SA and AH members is possible, many “comparisons end up comparing apples and oranges because students who go abroad are different from students who choose to stay in their home institutions” (Zaytseva et al., 2018, p. 3).

2.3.2 Alternatives to AH and SA designs

SA designs which do not use a control AH group to make comparisons can still provide valuable information about linguistic changes occurring at different stages of the learning process. Such designs usually follow two different approaches. The first approach, using longitudinal studies, employs a variety of techniques to gather data at multiple time points before, during and after SA. Serrano et al. (2012) analyze Spanish-speaking learners of English over an academic year at a British university, measuring changes in the same participant over a period of time before their SA experience and then over the same period including SA. Samples of oral

and written production are assessed at regular intervals in terms of fluency, syntactic complexity, lexical richness, and accuracy. A series of studies by Sasaki (2004; 2007; 2009; 2011; 2016), demonstrated some positive benefits that SA can have on the language learners' writing development. During observation periods ranging up to 3.5 years she looked at how Japanese L1 students' English language writing skills can change on visits varying from 6 weeks to 11 months, to English-speaking countries. More recent methods of data collection show how learner performance can vary according to linguistic environment before and after an L2 immersion experience. Marijuan and Sanz (2017) looked at how new data elicitation techniques such as online surveys, internet blogs and e-journals could extract useful information which answered questions related to changes in learners' motivation, identity, and intercultural competence. Comparisons can also be drawn between SA programmes of different lengths. For example, Avello and Lara (2014) looked the possible differential effects of Length of Stay (LoS), on L2 phonology. Two groups of Spanish/Catalan ESL learners of English, following 3-month and 6-month SA programmes respectively, performed a reading aloud task before and after SA in an attempt to detect significant differences in the learners' production as a function of LoS.

The second approach looks at how different levels of SA participant proficiency might affect learner outcomes. Duperron and Overstreet (2009) found that although short-term SA provided the most significant influence on students with weak language skills, generally results demonstrated SA experiences can benefit language learners of different proficiency levels in different ways.

2.4 SA Research: Aspects of language change

This section looks at some of the challenges in detecting subtle changes in language proficiency noting that different aspects of language tend to develop at different rates. It proposes that concentrating on one aspect of language, namely productive vocabulary knowledge, may offer the greatest chance of success. There is considerable evidence to support the view that productive vocabulary knowledge is strongly connected to other aspects of language proficiency and can be an important predictor of performance in a range of other skills.

2.4.1 Challenges in SA methodology.

There are a number of other important limitations with SA evaluation which go beyond the difficulty of arranging laboratory-like conditions or ensuring sufficient degrees of randomization in participation and research. Small samples and single case studies used in many research projects can result in low statistical power and contribute to the lack of reliability which can make generalizations more difficult. Many measurement tools and techniques presently in use lack sensitivity and are incapable of distinguishing minor changes in linguistic output. Sanz and Morales-Front (2018, p. 3) stated that “the coarse nature of the tasks implemented” in many experiments made them “unable to detect subtle changes, especially when learners are in the more advanced stages.”

Matters become more complicated with the fact that there is evidence that not every aspect of language will see the same level of impact from experience abroad. Some will show evidence of change within a short period while others will show little after a considerable time. Research into language fluency has produced evidence of significant positive changes. Serrano et al. (2012) analyzed the progress of 14 Spanish-speaking learners of English during a year spent abroad at a British university. They collected oral and written data during this period which were later analyzed in terms of fluency, syntactic complexity, lexical richness, and accuracy. The results of the statistical analyses indicated that while gains in oral performance were detectable within less than a month, improvement in written production seemed to be much slower.

As fluency skills seemed to undergo some improvement, phonological development seems to lag behind. Avello and Lara (2014) examined two groups of undergraduate Spanish/Catalan EFL learners following 3-month and 6-month SA programmes. Participants of each group were recorded performing a reading aloud task before and after their SA experience. A group of native speakers of English provided baseline data used to evaluate the learners’ phonological development. Results suggested no strong effect of SA and no significant differences in the learners’ production as a function of length of stay.

2.4.2 SA research and vocabulary change

There has been some difficulty in finding connections between an SA experience and its possible effect on certain aspects of language knowledge.

According to Zaytseva et al. (2018) empirical research has not yet demonstrated clear-cut findings of a link between SA and vocabulary improvement. However, it is nevertheless likely that the SA environment with its many opportunities for target language practice can enhance communication competence and speed up growth in vocabulary knowledge. A number of researchers (e.g., Lara, 2014; Pérez-Vidal, 2014; Zaytseva, 2016) provide evidence that vocabulary is among the top areas that improve the most after an immersion experience, well above reading, writing and grammar skills.

Lara's (2014) study explored the L2 oral development of a group of 47 adult EFL learners who participated in SA programmes. She compared the progress made by two groups of learners who went abroad for periods of 3 months and 6 months respectively. Results indicated that learners' oral fluency increased considerably during SA with lexical complexity moving toward more target-like use. Meanwhile there was little or no change observed in measures of syntactic complexity and accuracy. Zaytseva's (2016) study adopted a slightly different approach. It investigated the impact of two different consecutive learning contexts, formal instruction at home and a 3-month SA experience, on EFL vocabulary acquisition in oral and written production. 30 Catalan/Spanish advanced learners of English were assessed before and after each learning period using an oral interview and a written composition. These samples were analyzed using a series of lexical proficiency measures including fluency, density, diversity, sophistication and accuracy. Results revealed that SA is particularly beneficial for written productive vocabulary, and less so for oral, and that progress occurs especially in lexical fluency and diversity. With formal instruction, the study showed that there was a slight improvement in oral productive vocabulary lexical sophistication.

It does seem from past studies that there is some evidence to support the view that SA can have an impact on vocabulary knowledge. An often repeated observation from teachers is that learners carry around dictionaries and not grammar books especially when they travel abroad (Schmitt, 2010; Zaytseva et al., 2018 and many others). Even with the digital age and global communication technologies, online translating tools and dictionaries are the most frequently used applications that language learners turn to when using their mobile devices. In particular the use of tools such as Google Translate has widened as their perceived accuracy has increased (Ammade et al., 2023). More learners are viewing the use of mobile devices as an

important tool with which to enhance English language acquisition (Nuraeni et al., 2020). An early study on mobile phone use found that 45.2 % of students used dictionary applications and 32.9 % translation applications when studying a foreign language while only 5.5 % used their smartphones to access other forms of language learning application (Simon & Fell, 2012). A later study found that the use of digital devices to support vocabulary learning is widespread and often takes place outside the classroom (Brick & Cervi-Wilson, 2015).

Given the apparent importance attached to vocabulary it might make sense to concentrate on this particular aspect of language when we consider how language proficiency in general can be affected by SA. This could become more relevant if we attempt any meaningful assessment of the increasingly popular shorter-term SA programmes that we discussed in section 2.2.1. Established generalized tests like TOEIC and Eiken³ tests do not seem sensitive enough to detect changes in language proficiency over short periods of time (Drake, 1997). Therefore, there appears some need for a more sensitive way to measure language proficiency. Looking more closely at a single aspect of language change, namely vocabulary knowledge, may offer a solution.

2.4.3 Vocabulary knowledge and reading skills

One possible advantage of concentrating our measurement of SA language proficiency on vocabulary knowledge instead of attempting a more comprehensive assessment is that there seems to be a considerable amount of evidence linking it with other forms of language proficiency. For example research has shown that there is a strong relationship between vocabulary size and reading. Laufer (1992) explored the relationship between vocabulary size and academic text comprehension scores and attempted to find the degree of correlation. In her study 92 intermediate-level EFL learners took a total of four tests, two connected with each variable and two scores were calculated for each student: a reading score and a vocabulary size score. Highly significant correlations between 0.5 ($p < .0001$) and 0.75 ($p < .0001$) were found between both. Although Laufer had some concerns that other factors may also have contributed to reading comprehension scores, she concluded “that the results of the study support the claim that vocabulary size is a good predictor of the reading

³ EIKEN is an abbreviation of Jitsuyo Eigo Gino Kentei (Test in Practical English Proficiency), one of the most widely used English-language testing programmes in Japan.

comprehension level in foreign languages” (1992, p. 129). A study conducted by Qian (1999) had similar findings. He investigated the relationships between depth and breadth of vocabulary knowledge and reading comprehension using multivariate analyses. He examined the roles of depth and breadth of vocabulary knowledge in assessing the performance of ESL learners carrying out general academic reading comprehension tasks. The results support earlier studies in that scores on vocabulary size, depth of vocabulary knowledge, and reading comprehension were highly, and positively, correlated. Qian also found that scores on both vocabulary size and depth of vocabulary knowledge can be used to predict reading comprehension levels. Qian’s findings also recognized the importance of teaching depth of vocabulary knowledge as well as size in ESL education.

Nouri and Zerhouni (2016) examined the relationship between two dimensions of vocabulary knowledge, namely size and depth, and whether these two dimensions of vocabulary correlated with reading comprehension performance. Their research also empirically evaluated the tests used to measure these three constructs in the Moroccan EFL context. To this end, 32 freshmen specializing in telecommunication engineering at the National Institute of Posts and Telecommunication in Rabat, Morocco and taking English classes, were involved in the study. The instruments used include a) vocabulary size test, b) vocabulary depth test and c) reading comprehension test. The findings revealed a moderate correlation between size and depth of vocabulary knowledge, a significant strong correlation ($p < .01$) between depth and reading comprehension performance, but only a low correlation between vocabulary size and reading comprehension performance.

More evidence shows that learners tend to look for lexical information first when attempting understanding L2 texts. Laufer and Sim (1985) found that when learners attempted to interpret texts they relied on word meaning first, then on their subject knowledge and then least of all on syntax. These results suggest that the most important language skill threshold for reading comprehension was largely lexical. In a later study Laufer (1989) attempted to find how much lexical knowledge or threshold was necessary to enable adequate reading comprehension of an academic text. The results found that a learners’ vocabulary threshold size of 95% lexical coverage, that is knowledge of 95% of all word tokens was required to enable learners to read successfully and comfortably in an L2. Researchers have also tried to establish minimum vocabulary knowledge threshold sizes for reading similar texts.

Aizawa and Iso (2008) thought that that the probable minimum size was 6500 words out of 8000 words from the JACET 8000 wordlist (JACET, 2003) and Aizawa et al., (2009) found that there was no clear-cut boundary although they roughly estimated that a probable minimum vocabulary of around 5000 was needed to read an academic text reasonably well. These studies seem to indicate that while there is some doubt over the exact size of vocabulary knowledge needed for learners to become successful readers there seems no question over its importance.

2.4.4 The relationship between vocabulary knowledge and other skills

Some studies have attempted to link vocabulary knowledge with lexical inferencing success. Albrechtsen, Haastrup and Henriksen (2008) compared measures of vocabulary size and depth with several other measures which assessed learners' ability to use English. Using Danish ESL subjects they found that both L1 and L2 vocabulary size correlated with the ability to successfully make correct inferences, or guesses, of the meaning of unknown words in written text. These were significant correlations between 0.69 ($p < .0001$) and 0.82 ($p < .0001$) for the learners' L1 and between 0.48 ($p < .0001$) and 0.66 ($p < .0001$) for L2. Their L2 vocabulary sizes also correlated with L2 reading ability (between 0.73 and 0.80).

A comprehensive investigation of the relationship between vocabulary knowledge and language proficiency was carried out by Alderson (2005) as part of the development of a European project for the development of diagnostic language tests. The so-called DIALANG project offers separate tests for reading, writing, listening and grammatical structure and vocabulary for 14 European languages. The research team compared scores on a number of vocabulary tests with the scores from the other DIALANG tests. It seems that there was a strong correlation between vocabulary skills and other language skills when the results were considered. The vocabulary test battery correlated with reading at 0.64, listening from 0.61 to 0.65, writing from 0.7 to 0.79 and grammar at 0.64. Alderson concluded, "What the analysis would appear to show is that the size of one's vocabulary is relevant to one's performance on any language test, in other words that language ability is to quite a large extent a function of vocabulary use." (2005, p. 88).

Kiliç (2019) reported on empirical research that investigated the role of vocabulary knowledge in the writing and speaking performance of 54 B2 level Turkish learners of EFL. The measured aspects of vocabulary knowledge (productive

vocabulary size, receptive vocabulary size, and depth of vocabulary knowledge) were all found to correlate significantly with performance in writing and speaking (measured through the writing and speaking components of a proficiency test). Multiple regression analyses showed that vocabulary knowledge accounts for 26% of variance in writing performance and 17% of variance in speaking performance. Therefore, the study offers evidence that vocabulary knowledge is a significant predictor of performance in productive language skills.

In this section I have reported on evidence of connections between vocabulary knowledge and a range of other language skills, supporting the view that vocabulary knowledge measurement can be proxy for general language proficiency (Alderson, 2005). As has been shown in a number of studies, a range of L2 aspects including oral and written productive skills and reading and listening demonstrate a strong correlation with vocabulary knowledge. The conclusion that can be drawn from the research, such as that undertaken by the DIALANG project among many others, generally suggests that vocabulary knowledge can indeed be a useful predictor for other L2 skills.

2.5 Research on SA vocabulary acquisition

In the previous section (2.4) I have tried to establish the important role that vocabulary knowledge plays in overall language proficiency and have attempted to show how vocabulary skills correlate well with a number of other aspects of language learning. However, while an L2 environment is more likely to facilitate vocabulary acquisition and language development in general, it remains unclear precisely which particular areas of vocabulary knowledge itself undergo the greatest improvement. A look at some previous studies of L2 vocabulary development during SA might help us understand how changes can occur within a number of different aspects of vocabulary knowledge including receptive and productive skills.

2.5.1 Different aspects of vocabulary development

Sanz and Morales-Front (2018) presented a comprehensive summary of research on L2 vocabulary development during SA and considered a number of different aspects of vocabulary knowledge. This summary looked at past SA studies which considered a range of different dimensions of vocabulary knowledge. They described how vocabulary can be viewed in terms of its size (number of words

known) and sophistication (relative frequency of words), organization (connections made between words) and its accessibility, which is the speed and efficiency with which words can be recalled. There was also a distinction to be made between size and how well words are known (depth).

Until quite recently studies of vocabulary acquisition have lacked some consistency in their approach. Some of the measures employed have been less than precise and some of the sample groups studied have been difficult to compare in terms of size and participant profile (see sections 2.3.1 and 2.4.1). New research has begun to rely more on the use of empirical methods rather than impressionistic observations of vocabulary growth and looks more closely at the distinction between receptive and productive knowledge including the analysis of oral and written samples produced by participants. More recently, the importance of multiword units including formulaic language has also been recognized.

2.5.2 Some examples of vocabulary knowledge research

An example of SA vocabulary research which looks at productive knowledge includes a study by Milton and Meara (1995) who used a yes/no Eurocentre Vocabulary Size Test (EVST; Meara & Jones, 1990) to find that SA participants acquire vocabulary much faster than students who remained in their home countries (At Home or AH). They also discovered that those with low initial proficiency levels tended to make greater progress over time. However, this finding was not supported in a later study by Ife, Vives and Meara (2000) who found that learners from different proficiency groups in fact made similar gains. They also discovered from studying groups undergoing differing SA durations, that larger vocabulary gains were made with longer stays. Ife et al. (2000) used tests to measure vocabulary size (translation test) and organization (word association test) both of which were purportedly able to assess higher levels of proficiency. In an attempt to gain more reliable results Jimenez-Jimenez (2010) conducted a comparative study looking at participants undertaking SA along with an AH control group. Using similar measurement tools the study sought to establish the degree to which lexical items are known (depth of knowledge). The results show that SA benefited vocabulary gains both in terms of size and depth of knowledge compared with students remaining at home. Research by Dewey (2008) examined three groups in different learning contexts: AH learners receiving regular class instruction, AH learners receiving

immersion training and SA participants. It was found that the SA group scored more highly in all three vocabulary tests that were taken. The tests were designed to assess size and depth of vocabulary knowledge before and after the intervention event.

Sanz and Morales-Front in their 2018 summary pointed out that there has only been limited research of productive vocabulary knowledge partly due to the multi-foci of most SA studies. An example of such a multi-foci study would include that carried out by Laufer and Paribakht (1998) who used both receptive and productive vocabulary measures to examine the size and sophistication of vocabulary knowledge in different learning contexts. They found that learners receiving regular class instruction at home displayed a tendency to use less frequent words (greater lexical sophistication) than SA participants and also discovered that their receptive-productive vocabulary knowledge gap was smaller. Barquin (2012), using an approach which considers the Complexity, Accuracy and Fluency (CAF) of learners, examined two groups: one undergoing a 6-month period AH instruction and the other, 3-months of SA. She found that the SA learners achieved greater fluency in essay writing and lexical diversity. However, the lexical sophistication of both groups remained the same. Both of these multi-foci studies revealed some benefits of SA, particularly productive written knowledge, as well as receptive vocabulary size.

Looking more closely at oral productive vocabulary knowledge, some early studies tended to examine changes in grammatical as well as lexical competence. Ryan and Lafford (1992) looked at changes in Spanish verbs in a comparative study of two learner groups undergoing AH study and SA. Along with Dekeyser (1990) they found that SA learners were more accurate in their grammatical competence. Freed et al. (2004b) and Collentine (2004) also completed SA and AH comparative studies looking at lexical-grammatical competence. Both found that SA participants produced more language with increased informational richness. A study which looked more at lexical access with speech production rather than investigating grammatical changes was carried out by Segalowitz and Freed (2004). This comparative study found that SA participants' oral fluency improved significantly compared with AH contexts although there was much variation in word recognition and efficiency.

Lexical diversity is another issue worthy of consideration when investigating free oral productive vocabulary during SA. Unlike some previous research which mainly compared SA and AH contexts, Lara (2014) looked at how SA programmes

of different lengths affected the oral development of highly proficient learners. She found that it was difficult to detect changes in lexical diversity during a shorter 3-month stay but possible over a longer 6-month one (see section 2.4.2). Foster's (2009) findings further supported the development of lexical knowledge during SA by using more innovative measures with groups of SA and AH students. As well as finding that SA participants' speech production was more diverse there was evidence to support that with some individuals it actually approached native-speaker level.

Two more studies demonstrate that improvement in lexical accuracy can occur during speech production. Firstly, Llanes and Muñoz (2009) found that the number of lexical errors decreased over short (3-4 weeks) stays abroad. Although the study was originally designed to examine multiple variables, it found that lower proficiency participants in particular, produced more accurate and fluent speech. Secondly, Leonard and Shea (2017) discovered that SA learners tended to use more low frequency words in their active vocabulary after a 3-month SA period. However, there was no corresponding increase in lexical variety. With oral production the evidence seems to suggest that SA can have considerable impact on vocabulary knowledge. The greatest improvements seem to take place with semantic density, and lexical accuracy but less so with lexical diversity.

There seems to be a lack of research on both written and oral samples produced by the same participants in different learning contexts. Pérez-Vidal et al., (2012) investigated the effects of oral and written development on a group undergoing sequential AH formal instruction and SA learning experiences and revealed that while significant gains in lexical diversity were made, similar changes did not occur with oral speech production. These results are similar to those found by Freed et al., (2003) who discovered that written compositions produced by participants after an SA experience were longer and denser in lexical use than pre-SA compositions. Research by Serrano et al. (2012) found that improvements in oral fluency and lexical diversity seem to occur earlier than in written production over the period of a one-year SA experience.

Zaytseva's (2016) study has been previously mentioned in relation to the correlation between vocabulary skills and other aspects of language (see section 2.4.2). The same study also employs a comprehensive comparison of different kinds of vocabulary knowledge itself using both an oral interview and a written composition to enable their measurement. The interviews and compositions were

analyzed in terms of lexical fluency, diversity, density, sophistication and accuracy with results showing that SA participants' written compositions, in particular lexical fluency and diversity, improved more than oral interview performance. Zaytseva's research is unusual as she focusses closely on different forms of vocabulary knowledge as well demonstrating links with other changes in other, non-vocabulary, aspects of language proficiency.

Finally, some research focusses on possible changes in formulaic language which may occur during an SA experience. Studies by Möhle and Raupach (1987), Towell et al., (1996) and Regan (1998) reveal that SA learners were able to produce more fluent natural sounding language perhaps due to using more formulaic sequences. This is further supported by Foster (2009) who commented that many studies show that improvements to lexical organization, especially formulaic sequencing, can occur during SA. Furthermore Foster was able to demonstrate that immersion exposure to an L2 at an early age greatly helps to develop "native-likeness" (2009, p. 219). Foster's findings support Siyanova and Schmitt (2008) who found that an extended one-year SA period is more likely to contribute to more native-like idiomaticity. This section on formulaic language concludes that research carried out so far provides considerable evidence to support SA's role in developing more efficient and native-like formulaic language structures.

Considerable caution needs to be taken in generalizing some of the conclusions made about SA and vocabulary acquisition. With different kinds of participants, SA programmes and lengths, languages and measurement tools this caution seems to be warranted. A combination of different approaches is likely to provide the most complete picture of some of the complex processes taking place. Taking this into account, Dekeyser's (2014) ideas for a mixed-method approach seem to make sense as they consider how both the quantity and the quality of SA learners' interactions and other important factors might be a better way to predict language learning success. Briggs (2015) is a good example of such an approach as she investigated the relationship between out-of-class contact, conversation strategy and vocabulary in an SA context. Briggs' research is a reminder that it is important to consider factors beyond participant proficiency level and SA duration, in identifying effective SA programmes. The impact of different forms of accommodation, prevalence of Social Network System (SNS) use, specific

classroom language learning tasks and the way in which departure preparation might influence learning is as yet an under-researched aspects of the SA experience.

2.6 Measuring language change over short periods

Short-term SA programmes, or programmes which can be defined as six weeks or less in duration, have increased in popularity in spite of the fact that the measurement of any meaningful linguistic change over the limited time periods involved has proved no easy task. The results of previous research have been mixed. For example Drake's (1997) findings regarding changes in Japanese students' overall English proficiency levels during SA were inconclusive. Over the course of a 6-week programme in the United States students either moved up one level, stayed the same, or failed to maintain their original level when they took a general English test at the beginning and end of their stay. The test involved was the comprehensive G-STEP (Georgia State Test of English Proficiency) which separately examined reading, writing, listening and oral interview proficiency. No significant differences in student performance were detected on each occasion leading Drake to conclude that no current general test available, G-STEP or otherwise, was sensitive enough to measure six weeks of language learning. In my own informal study, I used TOEIC to assess changes in Japanese students' L2 proficiency during a 3-week SA programme in the United Kingdom (Caton, 2013). Again this general test detected no significant change in scores. There have been positive findings, too. Llanes and Muñoz (2009) examined learners' linguistic gains through oral fluency and accuracy measures as well as a listening comprehension task during a three-week SA experience. Their results revealed that an SA short stay does indeed produce significant gains on most measures and that proficiency level can strongly affect the intensity of learners' progress.

2.6.1 New approaches to measuring change

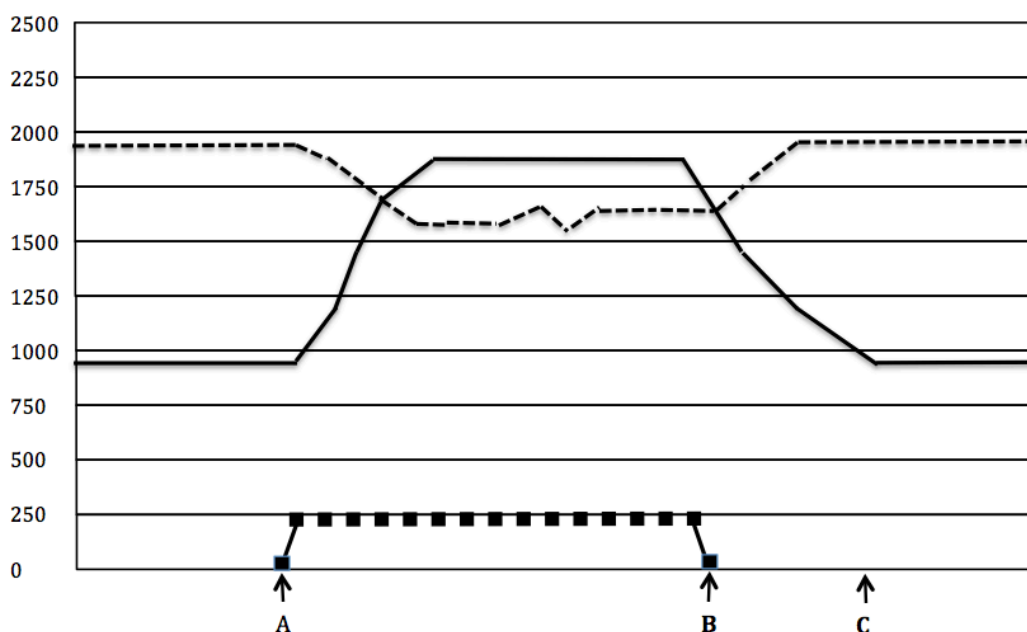
Such mixed results suggest that a new approach to measuring change is required. Within the relatively short time period exposed to an L2 environment during short-term SA, it can be assumed that the number of completely new words encountered and retained by learners will be limited thereby making any assessment of changes in word knowledge by conventional means of testing more difficult. Attempting to detect differences by taking existing known words and measuring their

shift towards how *well* they can become known, on the other hand, may offer a greater chance of success. One particular example of this change in the degree of knowledge that learners have about words is examining how words can move from being only passively known to becoming a full part of their active vocabulary knowledge. Furthermore there is some evidence that this movement can be detected within very short time periods as a study by Beaton et al., (1995) demonstrates. Their research showed that this particular aspect of vocabulary change, the shift from receptive to productive knowledge, can be detected within a matter of hours. In a single subject case study, using keywords to trigger or reactivate separate items, it was shown that an individual's knowledge of L2 vocabulary learned ten years previously could be rapidly recalled with a high degree of accuracy.

Meara (2005), provides another example of research into how this particular aspect of language change can occur over short periods. His paper described an attempt to elicit examples of spontaneous vocabulary reactivation as a result of a short period of immersion in an L2 environment. He starts with discussing the model

Figure 2.5

Meara's 2005 Model bilingual network (from Meara 2005, p.271)



The number of activated words in L1 and L2 is shown by the thin dashed and solid lines. Temporary stimulation of inactive words in the L2 beginning at A, raises the level of activation in the L1 words. When the temporary stimulation stops at point B, the initial equilibrium levels reinstate themselves, Point C.

shown in Figure 2.5 which looks at how an individual's L1 and L2 lexicons interact and assume a steady state where one becomes dominant over the other.

With the dominant lexicon a large proportion of the words are already in an activated state while with the L2 lexicon far fewer are activated, in other words, they remain dormant. In this way an L1 typically consists of vocabulary which is largely productive with relatively few words only known passively whereas the L2 usually is the reverse case with a larger passive vocabulary and a much smaller active one. Meara continues by comparing two separate cases where L1 and L2 dormant vocabulary are somehow activated. With L1 words the effects of this activation seem minimal but when dormant L2 words are activated more interesting effects occur. These effects are described as being "very extensive ripples of activation" which go on to affect a major proportion of an individual's L2 vocabulary (2005, p. 270). This change in dominance appears to be only temporary because once the activation stops the L1 reasserts itself and reverts back to the previous equilibrium. Meara goes on to compare this simple L1 and L2 bilingual model with a real world phenomenon affecting a number of L2 speakers. This phenomenon: "The Boulogne Ferry Effect" (Meara, 1999) is described as occurring in cases where a subject, who previously learned French as an L2 for example, travels to France to be once again immersed in an L2 environment and finds that their once dormant L2 spontaneously reactivates itself. An interesting point is that this reactivation seems to occur within very short periods, in fact within a matter of days. As most previous cases of this phenomenon consisted of informal anecdotal accounts and also because there was little evidence of any earlier research, Meara's (2005) paper described an experiment which sets about collecting real empirical data to overcome these shortcomings.

Meara (2005) explained that some disadvantages of using a traditional group study include the difficulty of justifying costs of sending subjects to a single overseas destination and also designing a suitable experiment while the nature of the phenomenon in question still remained unclear. Given these circumstances Meara instead decided to conduct a case study with a single cooperative subject. In Horst and Meara (1999) a computer-based measurement tool, *V_States*, was developed, which enables the user to identify, using a four-point scale from 's0' (I do not know this word) to 's3' (I'm certain I know this word), how well a word is known. For Meara's (2005) experiment an adapted version of *V_States* was used which required

the test subject to decide how confidently they could produce an English translation for each word out of a set of 244 Dutch stimulus words that had been learned years previously. The subject for the study was a L1 English speaker with dormant L2 Dutch who took the 15-minute test daily on 12 occasions before, during and after, a three-day trip to the Netherlands.

The results show that the subject did not know the translation for many of the L2 words in the period immediately before departure. However, when immersed in the new L2 environment it was found that there was a marked rise in the number of words in the s3 or “well-known” category and a corresponding fall in the number of words in the s0 or “unknown” group. On returning to a L1 home environment it was evident that many words remained activated although there was slight indication that some words were beginning to revert back to their dormant state. Meara concluded that environmental input is likely to play an important part in activating and maintaining vocabulary showing that the Boulogne Ferry effect is indeed a real phenomenon. Firstly, there was evidence that the subject’s active vocabulary actually more than tripled in size within two days of his arrival in the L2 environment and secondly, the study revealed that only a relatively small proportion, about 20% of the vocabulary tested, remained unaffected by the immersion experience.

In a search for a reliable means of measuring short-term language change there are a number of important points about Meara’s (2005) paper which should be made. Firstly, it successfully argues the case for single subject case studies, especially for cases such as this, which attempt to measure a phenomenon (the Boulogne Ferry effect) that is yet to be precisely defined, and which is therefore difficult to capture with existing research methods. The solution of using a single, carefully selected, cooperative test subject would help to overcome some of these concerns. At the same time problems which might affect a more traditional approach like using a group study are well described including the expense and difficulty of finding a sufficient number of experimental subjects with similar learner experience and proficiency profiles. However, it may be, that in certain cases, a group study may actually become more feasible particularly if the individual members within that group share a similar background. For example, if there is a mono-cultural group with the same L1, age, a similar proficiency level and who have experienced a similar learning environment there might be a reasonable expectation that most members share a similar initial state of vocabulary knowledge. This might not only

include the number of words known but also what many of those words are. An example of such a group might well be Japanese high school or university students who could satisfy many of these requirements. One important feature of Japan's MEXT is that it has created a rigid framework for the process of English education for most students during their compulsory schooling and at each stage of this there are a number of measurable parameters which can be used for assessment including word knowledge. A longitudinal experimental study which uses a similar homogeneous group might have a higher chance of producing valid data particularly when individual members of that group share a similar knowledge English base at the start. An intervention event in such a study like for example, an SA programme experience, would also expose group members to similar L2 environments and opportunities for vocabulary reactivation. In this way, Meara's study provides motivation for the research presented later in this thesis.

Secondly, Meara states in his paper that the full picture of whether vocabulary reactivation actually happens only emerges if a significant proportion of the total number of words a subject knows are tested. Although described as a boring and time-consuming process this is deemed necessary as many of the effects of reactivation are described as being subtle or otherwise difficult to detect. Using only small collections of words would make it more difficult to access many of these subtleties (2005, p. 272). However, recent developments in the field of corpus linguistics research might allow for clearer identification of small groups of words that may form a valid representation of a larger knowledge base. With Meara's single case study it was unusually fortunate that the subject had a written record of a large number of words that were once known and this allowed a study with convincing results to be made. In the case of a wider group study, however, a smaller carefully selected number of high frequency words might be used instead. These words, although not definitely known by every individual group member, would still have a high likelihood of being so implying that a study might still yield interesting results. Putting it another way, instead of starting an experiment with a single subject who possesses a definite active or passive knowledge of a relatively large number of words there would be a number of individual subjects with a very probable knowledge of relatively few. In both cases instances of reactivation would be more likely to occur during SA and could be detected through multiple testing. Since fewer words would be tested in a potential SA group study, administering the test on

multiple occasions would not present so many difficulties and a lesser degree of subject cooperation would be required.

A third point concerns Meara's explanation of the model and the graph (see Figure 2.5, p. 38). He describes the L1 reasserting itself after the stimulus activity has been completed and returning to its usual dominance while once again the L2 becomes dormant. In other words, both languages regain their former equilibrium state with correspondingly different levels of receptive and productive vocabulary knowledge. However, the experimental case study does not appear to completely agree with this earlier prediction as it shows reactivation still takes place, albeit it at a slower rate, after the subject returns to his L1 environment.

Fourthly, during the debriefing session at the end of the experiment, the single subject mentioned that repeated testing (on no less than ten occasions) of the same 245 words had had the effect of 'sensitizing' (Meara, 2005, p. 278) him. He thought that he was sometimes able to remember words, not as a result of spontaneous reactivation but because he encountered them in specific L2 environments. This made him wonder if only noticed the vocabulary of objects in his immediate environment or concepts he was directly experiencing only because he was often thinking about them beforehand. However, as the author points out, the fact that there is evidence that reactivation continues even after returning to the L2 environment suggests that word sensitization may not be such an important factor. Perhaps a future study could address this concern by using a choice of different randomly chosen words at each test event, decreasing the likelihood of sensitization taking place.

Finally, one of the most important points about the paper is that it acknowledges the graduated (incremental?) nature of word knowledge development. A idea first developed by Richards (1976) and refined by Singleton (1999), Gass and Selinker (2001) and Nation (2001) among others is that vocabulary knowledge should be conceptualized along a continuum where, at the one extreme are words which an individual has complete familiarity and mastery of (in Richards' terms, that would be knowledge of all seven aspects of his taxonomy of word knowledge) and at the other extreme are words with which they are unfamiliar, or which might be Zareva's "frontier words" (2012; 2014) - on the threshold of having some aspect of them acquired. Another way of expressing the same concept is that knowledge of a word is not an either/or state of affairs.

In summary, the following key points emerge from this section:

1. The evidence seems to show that conventional global tests, like G-STEP and TOEIC for example, lack sensitivity and are therefore unsuitable for measuring changes in learner proficiency over the short term.
2. Examining one aspect of language change, like vocabulary knowledge in this instance, seems to be able to provide more useful information about proficiency.
3. It is probably safe to assume that short-term SA does not allow time for a large number of words, encountered for the first time, to be learned by participants.
4. However, short-term SA does allow sufficient opportunity for many words that have been part of a learner's purely receptive knowledge to be reactivated and become part of their productive knowledge.
5. There is evidence which shows that it is possible to identify and track this shift from receptive to productive knowledge of words that are already known by learners.
6. In some cases the shift from receptive to productive knowledge can occur within a short period of time.

2.7 Comparisons of productive vocabulary tests

A number of points have emerged from the literature review so far.

1. Stakeholders involved in any SA programme require the need for some kind of evaluation of SA programmes. This is particularly true when it comes to measuring changes in language proficiency.
2. The process of language evaluation can be challenging. With SA there are multiple approaches and methodologies to contend with such as AH and SA

comparisons and longitudinal studies which compare before and after intervention language proficiency changes.

3. As well as applying an appropriate methodology there is also the question of what aspect of language to best evaluate. Relying on studies which try to measure multiple language aspects may not provide clear evidence of change.
4. There is evidence that vocabulary correlates well with a number of other measures and can act as a proxy for overall proficiency.
5. Vocabulary tests are available that address vocabulary size, sophistication, accessibility, depth of knowledge and that tap into productive or receptive knowledge. Appropriate test selection will be key to identifying changes during SA.
6. In the last section evidence from studies by Beaton et al. (1995) and Meara (2005) show that it may be possible to measure changes in productive vocabulary knowledge over a short time period. This made possible by using a test which can sense subtle changes with shifts from receptive vocabulary knowledge to productive.

2.7.1 Capturing vocabulary knowledge: Fitzpatrick and Clenton (2017)

I have so far established that some kind of productive vocabulary test is most likely to detect micro changes occurring during a short-term SA programme. The research described in this section will help us determine which one may be most suitable. A study by Fitzpatrick and Clenton (2017) offers a useful framework for doing this; they demonstrate that apparently simplistic vocabulary test scores tend to represent underlying complex sets of information and find that careful interpretation of the relationship between test scores and lexical competence is necessary for researchers and teachers better understand test application.

The authors point to the fact that a number of tests exploit the relationship between vocabulary knowledge and word frequency and that learners are more likely to acquire words according to the frequency of their occurrence in language use. Often depending on newly available pedagogical wordlists and corpora, the use of

such tests has become widespread. While tests are useful they may distract us from the fact that vocabulary knowledge can be also viewed as a multidimensional phenomenon with distinctions made between receptive and productive skills, partial and precise mastery and how words are organized in the mental lexicon (Cervetti et al., 2012; Stewart et al., 2012). This presents particular challenges for the development of productive vocabulary tests, for instance making distinctions between controlled and free productive knowledge testing and raising questions about representativeness and sample size. Fitzpatrick and Clenton (2017) set out to identify what exactly is measured by such tests and to reveal the ability of different tests to determine the quality of learners' knowledge of individual words and their overall vocabulary size.

They describe three different experiments. Firstly, language learners' performance on Lex30, a test of which I will talk much more later in the following section (2.8), is compared to their performance on an established test, the Language Frequency Profile (LFP; Laufer & Nation, 1995). Secondly, Lex30 is then compared with two newly designed tests; Brainstorm Frequency Profile (BFP) and G_Lex. Correlation analyses show that there are systematic differences in each tests' capacity to estimate information about the total number of words known by a learner and the thoroughness of individual word knowledge. The first experiment (N=80) shows no significant correlation in test performance between Lex30 and LFP suggesting that the difference in elicitation method used represent learners' lexicons in different ways, or different aspects of their lexicons. The LFP typically uses an essay-based writing task to generate a large body of text for analysis. Percentages of words contained within certain word frequency bands and words belonging to additional categories can then be calculated. Although the LFP is less restrictive than other productive vocabulary tests like the Productive Vocabulary Level Test (PVL; Laufer & Nation, 1999) in terms of the words that the test-taker is asked to provide, there remains an important disadvantage. Words elicited by a free writing task will be highly likely to contain a very high proportion of high frequency words which contribute little to the understanding of test-takers' breadth of productive vocabulary knowledge. An additional disadvantage is that the LFP requires considerable time both to administer and analyze the results. In order to see if the highly discursive nature of the task in LFP was the cause of this, a second experiment (N=80) comparing Lex30 to BFP, a nondiscursive equivalent to LFP, was carried out but

once again there was no correlation in test performance. A final experiment (N=100) compared Lex30 and G_Lex, a gap fill test with multiple activation events requiring contextual knowledge. The scores on both tests correlated significantly suggesting that the number of activation events and the quality of vocabulary knowledge accessed by the tests might have some kind of effect on the scores achieved.

These findings show how two dimensions of vocabulary knowledge, the total number of words known by a learner and the thoroughness of individual word knowledge, are represented differently depending on which test is used, highlighting a need for the careful interpretation of test scores. To understand learner test performance more clearly the authors formulate a model of vocabulary capture onto which the four test tasks can be mapped (see Figure 2.6, p. 47). Their two-dimensional model has a vertical dimension representing the quality of learners' word knowledge and a horizontal dimension relating to the proportion of lexical resource that the test task has the capacity to capture. Mapping the capture zone of different tests, each claiming to measure a similar construct, onto this two-dimensional model can help to explain the inconsistent correlations between test scores and can help teachers and researchers better understand their relationship with overall lexical competence.

2.7.2 Implications of Fitzpatrick and Clenton (2017)

The lesson that Fitzpatrick and Clenton's (2017) study delivers is that, although productive vocabulary knowledge tests like Lex30 generally categorize learner output according to word frequency, the way in which that output is generated will have a marked effect on how learner performance can be interpreted. The study illustrates the importance of elicitation technique in test design and some of the limitations of using a word frequency model as the sole means to assess vocabulary knowledge.

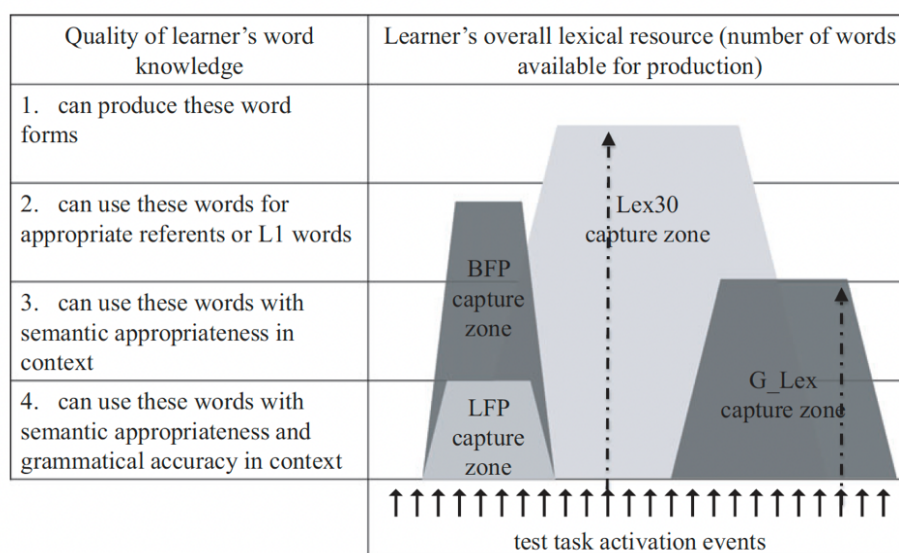
Normally making comparisons between different tests like Lex30 and LFP is not straightforward but in the study a good case is made for their similarity. They are both comprehensive tests measuring vocabulary knowledge as an independent construct and use the same model of vocabulary acquisition whereby the order in which words are acquired is associated with their frequency in general language use. Two more tests specifically designed for the investigation, BFP and G_Lex, are based on the same premise but vary, as with the first two, in the elicitation

techniques they use. By a process of experimental elimination the same learners' performance on Lex30 is not found to be comparable with that on LFP and BFP but does compare with their performance on G_Lex.

The Vocabulary Test Capture Model (VTCM) represents an effective way of exploring and understanding the differences between different productive vocabulary tests and shows that Lex30 is a task which is capable of measuring productive vocabulary knowledge in the broadest sense. Although there are a number of factors which should be considered when using the VTCM, the nature of its design means that Lex30 can include a high number of words with a capacity of being activated in some way (although not always necessarily accompanied with great knowledge or meaning) that causes learners to produce them (Fitzpatrick & Clenton, 2017, p. 20). In this the model might strengthen the arguments for Lex30's underlying validity by showing it is capable of making a far-ranging assessment of a learner's overall lexical resources.

Figure 2.6

Vocabulary test capture: Lex30, LFP, BFP, and G_Lex (From Fitzpatrick and Clenton 2017, p.19).



There are, as the study states, some further factors that might be considered (Fitzpatrick & Clenton's 2017, p. 19). Firstly, differences in learner proficiency are not discussed in detail but are often important when deciding which test to use. The LFP might prove challenging for low proficiency groups, for example, and Lex30

prove a more suitable choice. Perhaps there is some way this might be indicated in the model. The present horizontal arbitrary positioning of the capture zones for each test could be reconsidered with tests more suitable for lower proficiency groups placed on the left and those for higher towards the right (see Figure 2.6). Secondly, an even more effective way of differentiating between different qualities of learner word knowledge might be developed. Clearly a distinction between levels 2 and 3 should be made but this could be taken even further with the creation of more boundaries resulting in a more detailed and informative vertical scale.

For research on the effect of SA on learners' vocabulary knowledge Fitzpatrick and Clenton's (2017) study has some important lessons. Use of Lex30 can play a crucial role in experiments with SA participants as it has proved sensitive enough to detect short term changes in SA learners' knowledge. In the following section we will see how it has been trialed and evaluated in several studies, is suitable for low proficiency groups and is straightforward to administer and grade. However, what the VTCM does is to reveal some of Lex30's limitations particularly with assessing the quality of learner knowledge. Evidence points to the fact that a sizable proportion of words produced by test subjects in response to cues are only barely known and that the same subjects would be incapable of producing the same words in a sentence.

2.8 Lex30 explanation and validation

This part of the literature review will scrutinise the Lex30 test of productive vocabulary knowledge as a candidate test for gathering data in a longitudinal study. When investigating some of the complex changes occurring in learners' vocabulary knowledge during SA the first step will be to capture a representative sample of words produced at multiple timepoints before, during and after an SA experience. I will look at a number of key articles which describe how Lex30 was conceived, its methodology developed and how various validation and reliability studies undertaken by researchers continue to support its use. The limitations of this test will also be closely examined and a number of matters including learner proficiency, depth of vocabulary knowledge, scoring protocol and the effect of cultural and linguistic backgrounds, will be discussed. Examples of ways the test has helped further our understanding of how different aspects of vocabulary knowledge develop

over time are presented supporting the view that Lex30 can provide a valuable contribution to our understanding of lexical research.

2.8.1 Lex30 description

The original creators of Lex30, Meara and Fitzpatrick (2000), describe it as a tool which is capable of providing a straightforward assessment of the productive vocabulary knowledge of non-native speakers of English. Their efforts, which are described in their 2000 paper, represent a first step in introducing an innovative methodological tool while suggesting that it may be later developed into something more formal. The reasoning behind the test's creation is the recognition of the importance of learners' known vocabulary size, a factor which a number of studies have shown correlates closely with other language proficiency measures. Previous measurement of productive vocabulary size has been limited, relying initially on receptive measures and then extrapolating learners' active vocabulary knowledge from these. There has long been the assumption that learners' receptive knowledge is larger than their productive knowledge but the precise nature of the relationship between the two might not be so straightforward, underlining the potential usefulness of a new measurement tool.

Measuring productive vocabulary is problematic as words produced by the learner are likely to be context-specific. A small sample may not indicate the true and complete size or range of a learner's productive vocabulary knowledge and there might be difficulty with designing tasks which can capture a sufficiently large amount of vocabulary for subsequent extrapolation. Existing productive vocabulary tests can be divided into those which are *controlled* with subjects prompted to produce predetermined target words (Laufer & Nation, 1999) or *free* where subjects' spoken or written discourse is analyzed and a word frequency profile produced. An example of a *controlled* test is the Productive Vocabulary Levels Test (PVL; Laufer & Nation, 1999) has been previously mentioned in section 2.7.1. This uses target words supplied by the test subject to successfully complete sentences with sufficient context being provided to complete the task. An important criticism of the test is that the subject may also be well capable of providing an alternative, less frequently occurring word choice for some of the supplied sentences, possibly revealing a higher productive knowledge level than the test would indicate (Meara & Fitzpatrick, 2000; Walters, 2012). A *free* test, the Language Frequency Profile (LFP; Laufer &

Nation, 1995), has also been previously discussed (see section 2.7.1) and uses an essay-based writing task to generate a large body of text for analysis. The authors draw our attention to difficulties inherent with both types of test. Controlled productive tests, for example, can be effective when measuring small vocabulary sizes as a higher proportion of words which are potentially known can be tested but less effective with more proficient learners as this proportion decreases.

Figure 2.7

Lex30 sample test with data (from Fitzpatrick and Clenton, 2010, p. 554)

Cue	Responses
attack	heart, panic, surprise, unprovoked
board	snow, games, sports, surf
close	up, shave, shop, comfort
cloth	dish, wipe, dry, sink
dig	hole, shovel, digger, earth
dirty	clean, joke, bloke, wash
disease	ridden, plague, injection, cure
experience	familiar, innocent, practise, unprepared
fruit	apple, banana, salad, orange
furniture	warehouse, suite, wooden, leather
habit	bad, dirty, force, disgusting
hold	close, tight, grip, together
hope	pray, optimism, joy, faith
kick	boxing, flip, karate, fighting
map	reading, compass, grid, co-ordinates
obey	commands, obedience, demands, conform
pot	luck, plant, money, golf
potato	carrots, mash, spuds, sack
real	realistic, fake, imaginary, time
rest	sleep, up, wicked, peace
rice	cake, curry, fields, fried
science	scientist, fact, test, research
seat	belt, free, car, chair
spell	test, bound, magic, cast
substance	texture, material, test, feel
stupid	idiot, man, woman, thick
television	show, programme, star, set
tooth	brush, paste, decay, blue
trade	market, cards, games, money
window	cleaner, pane, glass, frame

Meara and Fitzpatrick describe the Lex30 as a task which involves word association. Subjects are presented with a list of stimulus or cue words which require them to produce up to four responses. These responses are not predetermined and have only limited constraints imposed by the chosen stimulus words. Figure 2.7 shows a sample Lex30 task with responses. The 30 words are taken from Nation's first 1000 wordlist (Nation, 1984) enabling the task to be completed across a wide range of proficiency levels. In addition, the stimulus words were selected using data from the Edinburgh Associative Thesaurus (Kiss et al., 1973) a database of word association responses. This enabled the selection of cue words (1) that tended not to elicit single dominant responses and (2) elicited responses at least half of which were not included in Nation's first 1000 list and could therefore be considered infrequent. In this way stimulus words like *black* or *dog* were avoided and those which chosen were more likely to generate a wider variety of associate words. Meara and Fitzpatrick (2000) describe their first experiment with Lex30 where 46 adult learners are chosen from across a range of proficiencies and L1 backgrounds. These learners are asked to write up to three words (four words in later versions) in response to 30 cue words verbally given at 30 second intervals thereby producing a maximum possible score of 90 (30 x 3). They also complete a standard yes/no vocabulary test (Eurocentres Vocabulary Size Test or EVST: Meara & Jones, 1990). For the purposes of scoring each of the responses are lemmatized with frequent derivational suffixes and affixes being counted as examples of a single base word. The text produced by each subject is then processed by computer to ascertain the frequency level of each word and a frequency profile produced. Words outside the 1000 most frequently occurring are awarded one point up to a maximum of 90 points.

Most of the words produced by each subject fall into Nation's first 1000 wordlist with about a third falling outside on average. Comparisons with the standard yes/no test show that students with a large receptive vocabulary also tend to produce a large number of infrequent words with Lex30. The high correlation seems to suggest that productive vocabulary knowledge size can be predicted by the size of learners' receptive knowledge. The results also show that receptive knowledge size increases in relation to productive size leading to a larger gap between the two as learners become more proficient.

These findings seem to suggest that Lex30 can be useful tool in measuring productive vocabulary knowledge. Its scores are sensitive to large differences in proficiency and relate closely to a test of receptive knowledge (EVST). The authors emphasize that the great advantage of the test is that it is easy to administer, taking only 15 minutes, and can easily be included as part of a larger test battery. The text that the test generates tends to be lexically rich with most words giving more useful information in comparison to previous methods. Meara and Fitzpatrick (2000) propose the idea that Lex30 can be used as a diagnostic tool by drawing attention to the fact that some test subjects' productive vocabulary size is not in direct proportion to their receptive vocabulary size. There perhaps might be some subjects who are not given a sufficient opportunity to develop their productive skills while overly concentrating on their receptive ones. In such cases Lex30 can be used to guide specific training regimes to take care of such an imbalance.

Meara and Fitzpatrick (2000) conclude that Lex30 has practical advantages over other currently used tests such as free productive (Laufer & Nation, 1995) and controlled productive (Laufer & Nation, 1999) tests and that it correlates highly with tests of receptive vocabulary. In addition it offers potential as a diagnostic tool which might be capable of detecting certain complexities in the vocabulary learning process.

2.8.2 Lex30: Possible limitations

Although Lex30 at first glance seems to offer a number of advantages there are a number of reasons to be cautious about its application. Firstly, the method of test delivery warrants further scrutiny. The paper's methodology section describes the administration of the test where subjects are provided with one one-word cue at a time and then given thirty seconds to write down up to three associate words. Other methods which have been described use computer based input. This is where cue words are presented on a computer screen and the participants are asked to type in their responses. Clenton (2006) describes an experiment which compares a standard form of Lex30 with an alternative version where subjects produce spoken responses which are then transcribed. There must be some question whether differences in the method of delivery of cue words or the way in which subjects record their responses has any effect on the number, frequency of occurrence and kinds of words that are generated.

The second point concerns the matter of proficiency level. When testing learners at very low levels there is a risk that they may not recognize the meaning of all of the cue words, precluding them from providing responses. This might be particularly important for young learners who are only likely to have receptive vocabulary of a limited size. Jiménez Catalán and Moreno Espinosa (2005) carried out a study using Lex30 with a group of 282 Spanish 10-year old school learners of EFL. The purpose was to ascertain whether the Lex30 tool could measure the productive vocabulary of young learners. They found their results were inconclusive and suggested that there were difficulties arising from the word frequency list used in the scoring procedure they followed. Alejo González and Piquer Piriz (2016) conducted a similar study with older secondary school students (16-year olds). Their study had similar aims and focused on assessing the validity and reliability of Lex30 when measuring the productive vocabulary of two groups of students (N=48). This time the results obtained were more positive indicating that Lex30 could be a more appropriate test when used with older and more experienced learners. These two studies seem to demonstrate that younger or lower proficiency learners may not be able to provide vocabulary samples of sufficient size with which to conduct a suitable analysis. Also certain groups of learners may not be familiar with many of the cue words even though they are chosen from Nation's first 1000 wordlist.

A third issue which may be problematical is the described procedure for test marking. There is the question of how should learners' responses in their L1, which also happen to be loan words used to varying degrees in English, be assessed. Common examples in Japanese might include the words *typhoon*, *judo* and *karate* and less common ones (but still familiar with a large section of the native speaker population) for instance *wasabi*, *mirin* and *sumo*. Perhaps the inclusion of such words in an established reference source such as a dictionary or even Wikipedia might make their use legitimate for the purposes of the test. As previously mentioned in section 2.3 there is an increasing awareness and acceptance of the diversity in the varieties of English that learners can produce. In this case Japanese learners are perhaps more likely to produce a greater number of words which may be less usual but nonetheless acceptable in normal language use. To this end Dewey (2012) points out that such learners should not be penalized for "innovative forms that are intelligible" (2012, p. 163). Loan words are not adequately explained in the paper's marking protocol and would be likely to have an important effect especially when

there are a considerable number of such words from test subjects' L1 which occur in English.

A final question which needs further thought concerns the selection of cue words. Although many of the criteria for cue selection are clearly described in Meara and Fitzpatrick (2000) there are two matters which should be further considered. The Edinburgh Associative Thesaurus (EAT; Kiss et al.,1973) which was compiled in the 1960s and the use of Nation's 1984 wordlist may not provide information that is entirely appropriate for modern learners. Modern alternatives now exist including the South Florida associative norms (Nelson et al.,1998) and the JACET 8000 Wordlist (JACET 2003). The former reflects more modern language usage and the latter is likely to have greater relevance for Japanese students who take the test. The practical advantages offered by Lex30 are considerable, but in applying the test, attention should be given to the methods of test delivery, subjects' proficiency level, marking procedures and allowing for changes in language use and cultural background.

2.8.3 Reviews of Lex30

Following Meara and Fitzpatrick (2000) some of Lex30's outstanding issues were assessed by Baba (2002) in her review of the test. She identified four main areas she felt warranted further discussion. The first concerns the nature of the responses that were produced in relation to the stimulus words. While understanding the fact that Lex30 uses a free word association task to gather data, non-native speakers tend to produce "idiosyncratic and unstable responses" (p. 68) to tasks compared to native speakers. She accepts that such behavior might still adequately reflect productive vocabulary knowledge although warns that some caution be exercised when interpreting results.

Secondly, Baba questions the construct validity of Lex30 and whether it is actually capable of assessing vocabulary knowledge without providing a context for words to appear in. She noticed that 65% of the words produced by subjects were frequent (that is, they were included within the first 1000 words) thereby showing that they were given credit for fewer than half the words that they produced. This, she concluded, may lead to some inaccuracies and greater difficulty in interpreting results.

Thirdly, Meara and Fitzpatrick used a split-half method to measure the test's internal consistency to show that any results produced were reliable and stable and to

demonstrate that infrequent words were not produced randomly. Baba concedes that the measure's correlation figure of 0.84 ($p < .01$, $N = 46$) is relatively high compared to similar tests but urges the use of further measures such test-retest studies or creating similar or parallel forms of the test to more firmly establish its reliability. The fourth point that Baba mentions concerns the avoidance of bias, especially with the selection of cue words, to ensure that there is no bias against groups from a "particular culture, gender, background knowledge or personality" (Baba, 2002, p. 70). Connected with this is the issue of fairness and the fact that the test only assesses written performance and not spoken. For example, certain test subjects (e.g., those who do not use Latin script) may well be competent L2 speakers but be less familiar with written word forms and therefore unable to correctly respond to the provided stimulus words.

Baba presents a concise assessment of the test, and her review appreciates its strengths, particularly the ease of administration and the fact that it seems to measure the construct that it is intended for. She supports Lex30's use, not as a replacement for existing productive vocabulary tests, but rather to augment them perhaps by detecting slightly different aspects of vocabulary knowledge. She mentions the comparatively low proportion of infrequent words produced (65%) but this still compares favourably to earlier productive vocabulary tests such as the Lexical Frequency Profile (LFP; Laufer and Nation, 1995).

Partly in response to Baba's 2002 paper and the attention that Lex30 received, Fitzpatrick and Meara (2004) conducted a new study further exploring the validity of the test. Underlining the fact that their previous 2000 paper only represented a tentative first step in the development of innovative measurement tool, this subsequent research investigated its reliability and validity using a test-retest study and two concurrent validity measures. As well as demonstrating that Lex30 is capable of producing valid and reliable results the paper suggests further improvement measures which go some way towards increasing our understanding of the complex nature of vocabulary knowledge. Lex30 is described as a easy to administer tool which generates lexically dense texts compared to other methods and correlates significantly with other vocabulary size measures.

Meara and Fitzpatrick (2004), in the first of three experiments, used 16 subjects from a range of proficiency levels to complete the same Lex30 test on two separate occasions with a 3-day gap to minimize any "practice effect" (Bachmann,

1990). A comparison of mean scores found no significant difference between scores at each test time. The authors note that while individuals' scores (i.e. the number of infrequent words produced) were very similar at test Times 1 and 2, the actual words they produced differed considerably. The experiment demonstrated that Lex30 has test-retest reliability, specifically addressing one of Baba's main concerns.

The other two studies reported in Fitzpatrick and Meara (2004) are two separate validity studies, the first a comparison between non-native speaker and native speaker norms and the second, a collateral test designed to distinguish between non native speakers of different language proficiencies. The first validity study compared the performance of two different subject groups, 46 adult English L1 speakers and 46 non-native speakers, on the same Lex30 test. The study found a significant difference between the mean scores of L1 and L2 speakers, but the separation of scores was not absolute, with an overlap between the scores of the most proficient L2 speakers and native speakers. The experiment shows that the test has some validity in so far as it is capable of distinguishing between native and non-native speakers.

The second validation study looks at collateral tests, which aim to distinguish between non-native speakers of different language proficiencies. It makes a comparison between Lex30 and other tests which purport to measure a similar construct. It uses two examples, firstly the Productive Vocabulary Level Test (PVL; Laufer & Nation, 1999), which evaluates productive vocabulary knowledge by requiring the subject to complete 18 partially given target words from five word frequency bands and secondly, a straightforward translation task. This requires subjects to translate 60 words from three separate word frequency bands from their L1 (Chinese) into English. Both tests, like Lex30, measure productive rather than receptive vocabulary and both focus on the use of frequency bands. The results showed that there was a strong correlation between the Productive Levels tests and the Translation task (0.843 $p < .01$) but a weaker one between them and Lex30, 0.504 (Productive Levels) and 0.651 (Translation task). Fitzpatrick and Meara explain that this weaker correlation occurs as the tests are measuring slightly different aspects of vocabulary knowledge.

As well as presenting the results of the experiments Fitzpatrick and Meara suggest some possible improvements to Lex30. Firstly, they note that there are still some technical issues with the test. During the test-retest experiment, frequency

profiles (but not the corpora itself) of the subjects were quite similar and, on average, native speakers tended to gain higher scores. However, a small number of highly proficient non-native speakers scored higher than some native speaker subjects. Use of more up-to-date frequency bands created with newer wordlists such as JACET 8000 (JACET 2003) is suggested to improve the test's accuracy and sensitivity. The second issue considers the basis of the construct of Lex30. The degree of knowledge that many subjects have about the associate words produced when they answer the test may vary widely from well-known to barely threshold level. With the Productive Levels test the knowledge requirement is more exacting: demanding knowledge of form, meaning and collocation and an understanding of context of the target words they are required to produce. This illustrates the complex nature of vocabulary knowledge and the fact that there are many degrees or depths with which particular words can be learned all of which are changing and interrelated. The different productive vocabulary tests that are available address different aspects of word knowledge as we saw in sections 2.7.1 and 2.7.2 with the Vocabulary Test Capture Model (VTCM; Fitzpatrick & Clenton, 2017).

Towards the end of the paper the authors consider the fundamental question: What exactly does Lex30 test? It is apparent that subjects find that producing response words is easier than with the Productive Levels test as they are not required to demonstrate such a deep level of vocabulary knowledge. Read (2000) distinguishes between two kinds or levels of productive vocabulary knowledge: recall and use. Perhaps a concern with Lex30 is that it does not have the capability to be able to distinguish whether a learner has sufficient knowledge of a word to be able to effectively "use" it. It is clear that the term "productive vocabulary knowledge" is misleadingly simple as it seems that this kind of knowledge is comprised of multiple forms. The only solution suggested is to develop a battery of tests, each measuring a slightly different form of knowledge, instead of depending on a single one. Lex30 fills an important gap in this battery.

2.9 Conclusion

In this literature review I have firstly tried to introduce the reader to the field of SA, its history and growth and its recent trends. Particularly in the case of Japan, there has been a rapid increase in the number of SA participants up until the COVID-19 pandemic era. I have also outlined some of the reasons why there has been a more

recent shift towards shorter-term SA programmes. Post-pandemic, the future of SA is less certain although there seems some evidence to suggest that there is already a slow return to pre-pandemic levels of participation. The pandemic has also seen the emergence of both online SA programmes, which have the potential to benefit a wider population of students, and innovative assessment methodology which tends to rely less on face-to-face and more on online interaction.

Secondly, I have emphasized the increasing need for the evaluation of SA programmes. Although many of the stakeholders involved require some level of justification for the time and resources expended, it is also clear that the process of evaluation can be challenging one. The reasons for this might relate to difficult questions about which basic methodology to follow such as in the comparison of SA with regular study or AH intensive courses or in the tracing changes in proficiency using a longitudinal study, for example. In addition, with any evaluation of language proficiency there is also the question of what to measure. With the use of generalized testing or with methodology which focuses more on multiple aspects of language change there may not always be a level of sensitivity which is enough to detect small changes in individual aspects of language.

Thirdly, I have presented evidence that prioritizes vocabulary knowledge and suggests that concentrating on this aspect may offer the best chance of success when trying to identify smaller changes taking place over a limited time scale. If we are to further investigate vocabulary it is necessary to examine the range of measurement tools that are available. Some distinction can be made with some of these by dividing them into two separate categories: those which test receptive vocabulary knowledge and those which test productive. Past research indicates that there is greater likelihood in detecting short-term changes by relying more on productive knowledge especially when taking in account the evidence provided by Meara (2005) and others.

Fourthly, as there a number of productive knowledge tests available I have attempted to make comparisons with some of them and select the most appropriate taking into account factors such as administration, participant proficiency level and the size of vocabulary sample it is capable of collecting. Before arriving at a final decision to use the Lex30 test I have reviewed some research which investigate its reliability, validity and some of its limitations.

In light of the research considered in this literature review, a decision was made to use the Lex30 test tool to elicit samples of vocabulary at multiple time points before, during and after an SA experience. This data will be used to address the following broad research question: What changes can be detected in L2 learners' productive vocabulary knowledge after an SA experience? In the first stage of this investigation we will address the questions: Do the total number of words produced by L1 Japanese learners in response to the Lex30 task change following an SA period, and if so, how? and do the number of infrequent words produced by L1 Japanese learners in response to the Lex30 task change following an SA period, and if so, how? As a preliminary step in this investigation, the following chapter will present a replication study, in order to determine whether Lex30 data from Japanese EFL learners studying in the UK mirrors that was obtained in two previous experiments conducted under similar conditions (Fitzpatrick, 2003 and Fitzpatrick & Clenton, 2010).

In later chapters of this thesis, I will consider some additional aspects of vocabulary use, namely collocational, orthographic and semantic grouping and to see if any changes occurring in each during SA, can be traced. At the beginning of each of these chapters will be a further literature review which will generate questions directly related to the aspect of vocabulary under discussion.

Chapter 3: Lex30: a replication study

Measuring changes in language ability as a result of short-term Study Abroad (SA) programmes is challenging. As noted in the previous chapter, there seem to be no tests available that are sensitive enough to cover short periods of language learning (Drake, 1997). Often the degree of language improvement (or otherwise) is so small that is nearly impossible to measure. It has also been suggested that an assessment method which focuses on single particular aspect of language and knowledge is more likely to reveal subtle changes taking place over a short duration than more general testing techniques.

It is possible that some methods of lexical analysis, including productive vocabulary tests like Lex30, are sensitive enough to pick up such small changes in the overall language competence of short term SA programme participants. This replication chapter will first explore some of the findings of Fitzpatrick and Clenton's (2010) study which measures changes in productive vocabulary knowledge over a short period. The chapter will also look at a similar study by Fitzpatrick (2003) carried out as part of her thesis research. It will attempt to answer the following two research questions:

- 1) To what extent can data from a Lex30 test administered immediately before and after an SA period detect changes in productive vocabulary knowledge?
- 2) Do the findings from this study align with those reported in Fitzpatrick and Clenton (2010) and Fitzpatrick (2003)?

In particular it will examine the results of longitudinal studies carried out with SA participants and see how alternative methods of scoring the Lex30 can give very different results. Having considered these studies, the chapter will then report on a replication experiment conducted by the author, using data from Japanese L1 students participating in SA programmes in the UK. The findings from that study will be compared with those from the studies mentioned above.

3.1 The Lex30 productive vocabulary test

The Lex30, originally developed by Meara and Fitzpatrick (2000) has already been described in some detail (see section 2.8). It was created in response to the need to find an alternative to controlled productive tests (e.g., Laufer & Nation, 1999) which I have explained were found to be useful but problematical. It has been suggested that controlled tests can be effective with low proficiency subjects with a limited vocabulary size since a relatively high proportion of their total knowledge can be tested. For subjects with much larger vocabularies it becomes difficult to extrapolate their entire knowledge using a test with only a limited number of questions. We have also seen that free productive vocabulary tests (Laufer & Nation, 1995) have their problems. In most cases they use a context limited essay-type assessment style to encourage subjects to display as much vocabulary knowledge as possible. Most words elicited by such essay-like tasks are likely to be frequently used and a huge amount of text seems to be required to generate a sufficient number of infrequent words that are needed for assessment. Free tests can work but in practice they have proved impractical as they take a considerable time to administer and score.

Lex30 has attracted much interest from teachers and researchers as a practical testing measure and it has shown several advantages over the productive tests described earlier. It seems simple to administer, complete and score and is perceived to have high “face validity” (Fitzpatrick & Clenton 2010, p. 538). Despite these seeming advantages concerns certainly exist, first highlighted by Baba (2002) and then Clenton (2005; 2008). Fitzpatrick and Meara (2004) have addressed many of Lex30’s shortcomings by conducting experiments on test reliability, concurrent validity using native speaker comparisons and concurrent validity using collateral test measures. Walters (2012) has replicated some of these previous investigations with different populations and has gone on to look at the test’s ability to distinguish between different proficiency levels. She concluded that the Lex30 can be a “useful and valid test” (2012, p. 184). For this replication study we shall first examine Fitzpatrick and Clenton’s (2010) paper paying particular attention to the description of a longitudinal experiment carried out as part of a series of comprehensive validation studies.

3.2 Measuring changes in vocabulary knowledge: Previous studies

3.2.1 Fitzpatrick and Clenton (2010)

Fitzpatrick and Clenton's paper assesses many aspects of the performance of the Lex30 vocabulary test. They build on previous findings (Meara & Fitzpatrick, 2000; Fitzpatrick & Meara, 2004) and take a further look at the test's reliability and construct validity. The aim of their research is not to argue for, or against, the validity of the test per se but to thoroughly explore its potential and to identify its limitations. The overall usefulness of the test is considered by structuring the paper around a series of issues previously raised by other researchers (Baba, 2002; Jiménez Catalán & Moreno Espinosa, 2005). The authors also consider how the Lex30 test fits in with Bachman and Palmer's "usefulness" formula (1996, p. 18) looking at the elements of any language test can be thought as universally important.

The paper describes the Lex30 test's ability to elicit lexically rich text in an economical way comparing it favourably with other productive knowledge measures including controlled productive knowledge tests (Laufer & Nation, 1999) and the Lexical Frequency Profile (LFP; Laufer & Nation, 1995). It describes the test as being able to elicit a wide range of vocabulary from different conceptual fields using a single word association stimulus. It is argued that the careful selection of these "cue" words minimizes the effort needed for their activation and maximizes the range of potential responses.

Three main experimental areas are covered. Firstly, the reliability of the test is looked at using a test-retest study where the test is administered on two occasions with an interval of a week, a parallel test forms experiment and an internal consistency measure. In the test-retest study test scores correlated significantly supporting Fitzpatrick and Meara's (2004) findings that it has a high degree of reliability. Parallel forms of the test were also found to correlate well with each other and were not significantly different and a calculation of Cronbach's alpha demonstrated an acceptable internal consistency. Secondly, the paper looks at the test's construct validity trying to determine whether the test reflects vocabulary improvement over time by administering it to the same group of L2 learners over an interval of 6 weeks during which the learners participated in a language improvement class. I shall return to this particular experiment in due course as it forms the focus for our particular replication study. Two other issues concerned with construct validity are explored: whether the test compares favourably with other similar tests

purported to measure similar construct and a comparison of spoken and written test performance. Although the Lex30 correlates significantly with other tests the degree of correlation seems relatively low and the correlation between the two test delivery forms also brought mixed results.

Returning to the second experiment which looks at Fitzpatrick and Clenton’s investigation into vocabulary change after a 6-week learning period, this connects with our research question (1) which asks: To what extent can data from a Lex30 test administered immediately before and after an SA period detect changes in productive vocabulary knowledge? (Fitzpatrick & Clenton, 2010, p. 543). Prompted by comments from Baba (2002) calling for more evidence for the validity of Lex30 and following observations from Bachman (1990) about gathering such information by comparing learners of different language proficiencies, it was decided to compare learner data with data from the same learners after a language learning intervention period when it might be reasonably expected that language proficiency had improved.

The longitudinal experiment is conducted using Lex30 to obtain criterion-related evidence on the validity of the test. It is designed to detect vocabulary knowledge increase over a six-week period with a group of 40 L1 Japanese pre-intermediate students attending English improvement classes. The Lex30 test was administered on two occasions: at the beginning and at the end of the six-week “language intervention” period. All the participants’ responses to the 30 cue words were lemmatized according to strict criteria (from Bauer & Nation, 1993 as described in Meara and Fitzpatrick’s 2000 paper. See section 2.8.1 in the previous chapter). One point was awarded for every low-frequency word produced, with “low-frequency” being defined as not being in the 1000 most frequently occurring English words. For this study the JACET 8000 word list (JACET 2003) was used.

Table 3.1

Fitzpatrick and Clenton 2010: Longitudinal study score data

	N	Min.	Max.	Mean	SD
Test Time 1	40	9	42	24	8.514
Test Time 2	40	7	48	29	9.084

The descriptive statistics are shown in Table 3.1. The difference between the two means (24 at test Time 1 and 29 at test Time 2) was found to be significant ($t = 4.825, p < .0001$). The two sets of scores correlated at 0.809 ($p < .01$). The authors conclude that the increase in scores shown here between test Time 1 and test Time 2 is evidence that the Lex30 test is capable of detecting a change, in this case an improvement, in learners' productive vocabulary ability. It was suggested that this significant improvement in scores was also not likely to be due to any practice effect as made evident in the reliability test-retest study they reported elsewhere in the paper.

3.2.2 Fitzpatrick's 2003 Thesis

Further building on her work with Paul Meara on the introduction of the Lex30 productive vocabulary test (Meara & Fitzpatrick, 2000) Fitzpatrick explores how productive vocabulary can be elicited and measured by using word association techniques and word frequency lists. For her PhD thesis study she looks at how the Lex30 test can collect useful data in an efficient way, thereby avoiding many of the problems that had plagued previous attempts at designing similar tests (e.g., Laufer & Nation, 1995; 1999). She describes the Lex30 as a test which uses a word association technique to allow subjects to produce a small corpus of words which is representative of their total productive lexicon. The absence of predetermined target words and narrow context constraints by a narrow context encourage subjects to elicit content words across a wider range of frequency bands than might otherwise be the case.

Much of her research looks at the painstaking development process of the Lex30, first looking at an early version of the test, the Lex100, which was carefully modified and refined into its present shape. Particular attention is paid to cue or stimulus word selection by using the Edinburgh Associative Thesaurus (Kiss et al., 1973) mentioned in Fitzpatrick (2003, p. 115), a database of word association norms, listing response words, and the frequency with which they occur, for 8,400 stimulus words. Then the process of lemmatization is examined using the formal set of criteria from Bauer and Nation's "Word Family" lists (1993). Once the test has been shown to work relatively smoothly in practice, Fitzpatrick takes us through several more stages looking at score consistency, native speaker comparisons and longitudinal studies generally looking at reliability and validity of the Lex30 test, and concludes

that the test has significant potential as a measurement tool. She does caution us to be aware of concerns about its accuracy and sensitivity and this will become particularly evident when we take a look at longitudinal test validation studies.

3.2.3 Fitzpatrick's 2003 longitudinal study

As part of her thesis Fitzpatrick reports a study which is highly relevant to the research we report later in this chapter. Its purpose was to see whether Lex30 tests taken at the beginning and end of a study period could detect changes in learners' language proficiency. As well as taking the Lex30 test, subjects also took the receptive Eurocentres Vocabulary Size Test (EVST; Meara & Jones, 1990) to compare any changes in their productive performance with changes in their receptive lexicon. The two study groups differ in that the first is in Britain for a period of 4 weeks while the second, although in the country for a year-long exchange programme, was tested before and after a 5-month period. For our replication study and for the purposes of a comparison with the experiment carried out by Fitzpatrick and Clenton (2010) we shall only look at the first group.

Fitzpatrick (2003) looked at 19 L1 Japanese undergraduate students participating in a 4-week intensive English language course at a university in the UK. Their age was between 19 and 23. During the course students received a minimum of three hours English language instruction per day while staying with local English-speaking host families. During the evenings and at mealtimes the host families were required to provide as many opportunities as possible for students to speak English. In order to maximize exposure to English only one Japanese student was permitted to stay with each host family. Initial placement tests found that language proficiency levels ranged from elementary to intermediate.

The subjects took the computer version of the Lex30 test on day 1 of their programme. At that time they had already been in Britain for 2 days and had spent a weekend with their host family. The subjects were also tested 24 days later during the last week of their programme. The test required them to type in as many responses as possible (up to a maximum of 4) for each cue word provided. The Lex30 scores were calculated according to a percentage method (Fitzpatrick, 2003, pp. 148-151). This means that each participant's Lex30 score represents the number of infrequent words they produce as a percentage of the total number of words produced. The subjects all took the Lex30 and the EVST together at both test times.

For the purposes of this replication study we are only interested in the results of the Lex30 test. Table 3.2 shows the longitudinal study score data. The difference in the mean Lex30 scores between test Time 1 and 2 was not significant. In other words, the Lex30 scores remained relatively stable over the 4-week period. The t value was: $t = 1.29, p = .213$.

Table 3.2

Fitzpatrick (2003): Longitudinal study score data

	N	Mean	SD
Test Time 1	19	22	6.97
Test Time 2	19	20	6.78

The individual Lex30 performances at Test times 1 and 2 show a significant correlation 0.636 ($p < .01$) between them. On the scatter graph (Fitzpatrick, 2003, p.189) and shown together with data from the replication experiment in Figure 3.2, p. 70) we can see the majority of subjects are placed above the line, indicating that they scored higher on Test 1 than at test Time 2. A summary of the results suggests that the number of infrequent words in the subjects' productive lexicons has not increased over the study period or perhaps that the Lex30 test is not sensitive or sophisticated enough to pick up any increases over a short 4-week period.

3.2.4 Towards a replication experiment

Both studies that have been described above purport to measure a similar construct: detecting changes in the productive vocabulary performance of students attending short-term SA programmes. The results from each are very different and this cannot be easily explained. A real difference in test performance is likely to be a factor but other influences may be at work, too. Differences in scoring procedures, test protocols and even learning environment may also play an important role. The following replication experiment will try to follow these earlier longitudinal studies and take account of some of these influences. First, it will essentially ask the same research questions as indicated at the beginning of this chapter (see p.60):

- 1) To what extent can data from a Lex30 test administered immediately before and after an SA period detect changes in productive vocabulary knowledge?
- 2) Do the findings from this study align with those reported in Fitzpatrick and Clenton (2010) and Fitzpatrick (2003)?

3.3 Methodology

The study collected data from Japanese L1 university students, before and after short-term SA visits. The participants were 38 female students aged between 18 and 21 years old attending three separate courses at two universities in Fukuoka, Japan. The course included SA visits to Canada, UK and USA respectively. Students completed a Lex30 test in the days before departure, and another on completion of the SA visit. The English proficiency level for the average participant was estimated to be advanced beginner level or CEFR (Common European Framework of Reference for Languages) Upper A1. The CEFR is widely accepted as being the global standard for grading an individual's language proficiency and the Upper A1 level is seen as roughly equivalent of a TOEIC score of between 230 – 280. Table 3.3 shows the background profile of the students and pre-test and post-test times. Ethical approval was obtained from the relevant university committees, participants received information sheet about the Lex30 task and completed a consent form.

Table 3.3

Replication Study (2016): Group participants

N=38	University department	SA location	Pre-Test days before departure	Post-test days after return
17	Nakamura: Career Dev	Vancouver, Canada	8 days	0 days
19	Nakamura: Nutrition	Canterbury, UK	3 days	0 days
2	Fukuoka University: Law	Hawaii, USA	5 days	3 days

In order to obtain a sample of reasonable size for the experiment three separate groups were used. All students spent 17 days in total in their respective study abroad countries staying with host families and undertaking a similar educational programme. They were given many opportunities to use their English and were encouraged to interact closely with their host families. There was a

maximum of one Japanese student permitted to stay with each family to ensure maximum English language exposure. During weekday mornings students attended English classes (10 days in total) at a language school and in the afternoon and at weekends they had an extensive programme of sightseeing and cultural activities. The three programmes offered an equivalent educational experience and there appeared to be no major difference between them.

Following Fitzpatrick and Clenton (2010), Lex30 was completed using a paper-and-pen method. It was felt that students would be more comfortable using this more familiar style of test administration than the computer version used by Fitzpatrick 2003. The pre-test was conducted eight days before departure for students going to Vancouver, three days before for students going to Canterbury and five days before for students going to Hawaii. The post-test for the Vancouver and Canterbury students was carried out at the language school just before their return to Japan while Hawaii students completed their test within three days after their return. The test was administered by three university staff members who conducted orientation classes and accompanied students on their programmes. Students were given a time limit of 15 minutes to complete the test on each occasion. The slight discrepancy in timings of pre- and post-tests were to accommodate slightly different school schedules.

After the tests were completed they were processed according to protocol as followed by Fitzpatrick (2003) and Fitzpatrick and Clenton (2010). All responses were individually lemmatized so that inflectional suffixes (plural forms, past tenses, comparatives) and frequent regular derivational affixes (-able, -ly) were counted as base forms of these words. The full criteria used by Meara and Fitzpatrick corresponds to levels 2 and 3 of Bauer and Nation's 'word families' (Bauer & Nation, 1993). Other protocols followed including awarding '0' points for proper names, numbers, Japanese words and acronyms from the corpus created by each student. For a full list of protocols covered please see Appendix 1. The JACET 8000 wordlist (JACET, 2003) was used to identify the 1000 most frequently occurring words. Each word in a test subject's individual corpus that did not appear on the list of 1000 most frequently occurring words was awarded one point. This total number of infrequently occurring words constituted a final raw Lex30 score.

3.4 Results

The descriptive statistics are shown in Table 3.4. The difference between the two means (18.24 at test Time 1 and 26.16 at Test time 2) was found to be significant ($t(19) = 6.8854, p < .001, d = 1.117$). Since the d is higher than 0.8 this indicates a large effect size. The two sets of scores correlated at 0.747 ($p < .001$). The increase in scores shown here between Test time 1 and test Time 2 seems to be

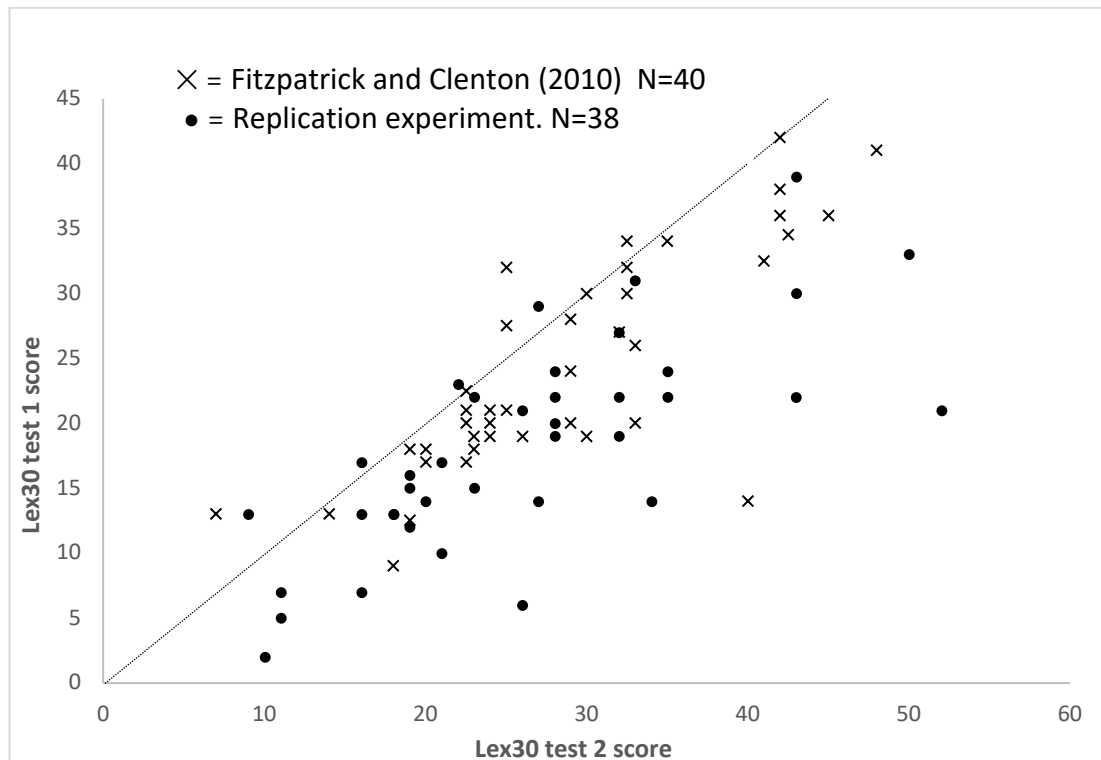
Table 3.4

Replication (2016): Lex30 scores before and after SA

	N	Min.	Max.	Mean	SD
Test Time 1	38	5	39	18.24	8.159
Test Time 2	38	9	50	26.16	10.666

Figure 3.1

Relationship between pre- and post-SA Lex30 tests: Replication v. Fitzpatrick and Clenton (2010, p.544)



evidence that the Lex30 test is capable of detecting a change, in this case an improvement, in learners' productive vocabulary ability. We can interpret this significant increase in scores as being a result of the participants attending an English SA programme.

We can conclude that, in this case, the Lex30 seems to be sensitive to improvements in learners' language ability. The individual Lex30 performances at test Times 1 and 2 for the replication study and for Fitzpatrick and Clenton's (2010) study are both illustrated on the scatter graph in Figure 3.1. The graph indicates the relationship between the scores at the two test times for the two studies. The line on the graph is where subjects who scored the same at both test times are plotted. The great majority of subjects for both groups are placed below the line showing they scored higher at test Time 2 than Test 1.

3.5 Discussion

A comparison of our replication experiment with Fitzpatrick and Clenton's (2010) findings shows many similarities as can be seen in Figure 3.1. In answer to the first research question whether data from a Lex30 test administered immediately before and after an SA period detect changes in productive vocabulary knowledge, the answer would most certainly be "yes". With the second question of whether the findings from the replication study align with those reported in Fitzpatrick and Clenton (2010) and Fitzpatrick (2003) then the answer would be affirmative for the first study but less than certain for the second.

Comparing Fitzpatrick and Clenton (2010) with the replication experiment the minimum and maximum Lex30 scores in both studies are very close for both the pre and post tests. For example the maximum score in the replication post-test was 50 compared to 48 for Fitzpatrick and Clenton (2010) while the minimum score was 7 compared to 9. The spread of scores were also similar as shown by the figures for standard deviation and Figure 3.1. In both studies the difference between pre and post test means was significant although the *t* value was slightly higher with the replication. The correlation of test Time 1 and test Time 2 scores is also close with figures of 0.809 ($p < .0001$) for Fitzpatrick and Clenton's study and 0.747 ($p < .0001$) for the replication.

In order to make a meaningful comparison with Fitzpatrick's 2003 longitudinal study it was necessary to convert the raw scores (infrequently occurring

words from JACET 1K+) gained in the replication study into a percentage form (the number of infrequent words produced divided by the total number of words produced as described in Fitzpatrick, 2003, p. 148). Very different scores were obtained with the percentage scores compared with raw Lex30 scores. In both the replication study (see Table 3.4) and Fitzpatrick (2003) the number of infrequent words produced by most participants increased from test Time 1 to Test time 2. However, the total

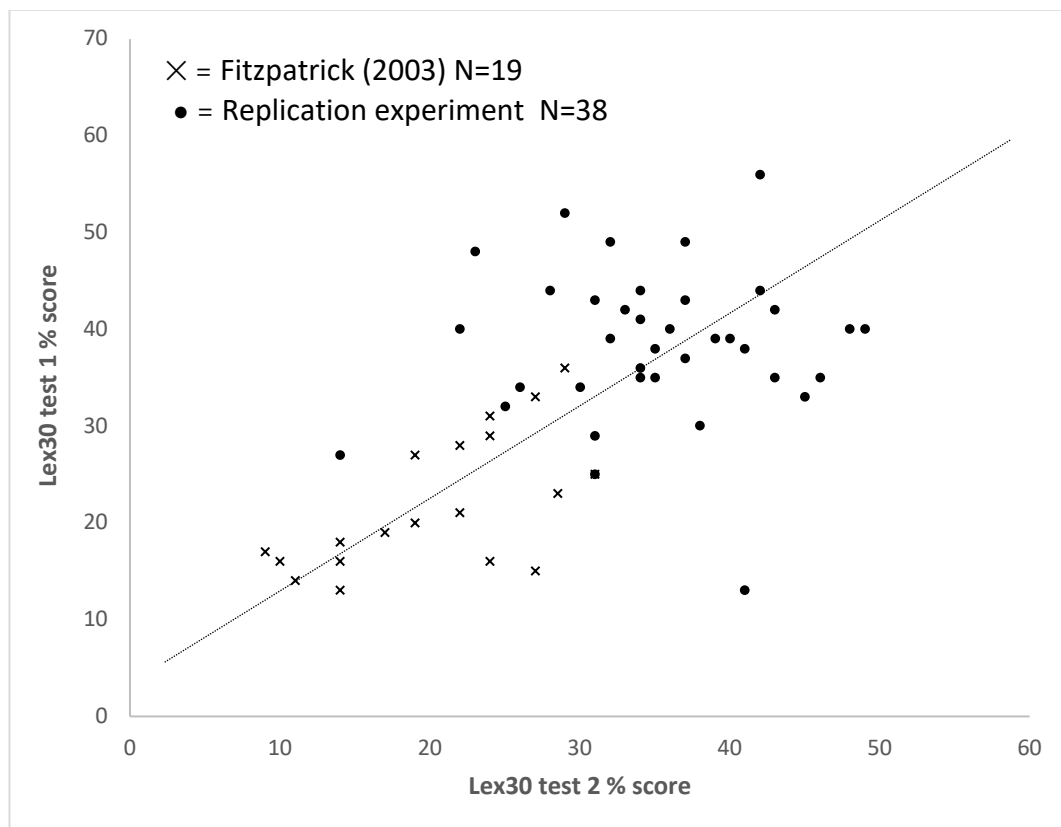
Table 3.5

Replication study (2016): Lex30 percentage scores before and after SA

	N	Min.	Max.	Mean	SD
Test Time 1	38	13	60	39	8.580
Test Time 2	38	14	49	35	7.680

Figure 3.2

Relationship between pre- and post-SA Lex30 percentage scores: Replication v. Fitzpatrick (2003, p.188)



number of responses increased over the same period at a much greater rate. The result of this is that the percentage of infrequent words as a proportion of the total number of words produced by each participant actually decreased.

The difference in the mean Lex30 scores between test Time 1 and 2 was not significant. As was the case with Fitzpatrick (2003) Lex30 scores remained relatively stable over the 17 day study abroad period. The t value was: $t = 2.079$ ($p = .045$). The individual Lex30 performances at test Times 1 and 2 show a moderate correlation 0.406 ($p = .406$) between them. The individual Lex30 percentage scores for the replication study and for Fitzpatrick's (2003) study at test Times 1 and 2 are both illustrated on the scatter graph Figure 3.2. The line on the graph is where subjects who scored the same percentage score at both test times are plotted. The data points for Fitzpatrick (2003) and the replication study display a similar pattern with roughly the same number of points apparent on both sides of the line.

The graph illustrates a weak correlation between the scores at the two test times of 0.139 ($p = .406$). This can be compared to a significant correlation 0.636 ($p < .01$) in Fitzpatrick's data (2003, p.188). There were a number of factors in the design and administration of the Lex30 test that might have had an influence on the eventual outcome. These issues will each be dealt with below.

3.5.1 Participants

One consideration is that instead of three groups studying different courses at two universities perhaps it would better to have one homogenous group which can be better controlled. There was some variation in the arrangements for test administration during the experiment because the groups were meeting and preparing for their overseas trips at different times. Whether or not every participants' experience was entirely similar may perhaps be of some concern given the differences in study environment in three different countries. However, it was considered that having a sample size of 38 students afforded more confidence in our results than a smaller sample from a single course and was close to the 40 subjects participating in Fitzpatrick and Clenton's study.

3.5.2 Participation period

The length of short-term SA programmes has gradually shortened in recent years due to academic, economic and employment reasons. The 17 days spent by

subjects on their short-term programme was considerably shorter than the four weeks and six weeks of the two studies that our experiment tried to replicate. There is evidence that the longer students spend studying abroad then the greater their gains in language proficiency. Milton and Meara (1995) found that SA students' vocabularies grew four times as fast compared to at-home learners and Llanes and Muñoz (2009) also correlated fluency gains with length of stay. Conversely, others argue that length of stay is less important than quality and quantity of contact with the target language while abroad (Bardovi-Harlig & Bastos, 2011). Although the SA period was short – a period of 17 days – significant gains were indeed found in raw scores, though not in percentage scores.

3.5.3 Proficiency level

As previously mentioned the proficiency level of each SA participant was estimated using the author's own judgment based on individual classroom performance. Generally there was some variation in proficiency levels as the gap between the highest and lowest Lex30 scores (5 and 39 for Time 1 and 9 and 50 for Time 2) seems to indicate. For practical reasons this study was not able to use the Eurocentres Vocabulary Size Test (EVST; Meara & Jones, 1990) in a similar way to Fitzpatrick (2003) as an additional measure.

It is unclear what kind of effect a better knowledge of the proficiency level of participants involved in the experiment might have affected the results. The experiment design was such that it was not possible to trace the degree of change of Lex30 scores (or in fact for any other proficiency measure) for individual test participants although this is certainly an aspect that could be considered in a future study.

3.5.4 Timing of pre and post tests

Care was taken with the timing of both pre and post tests. If the pre-tests are carried out too early then perhaps students would have further uncontrolled opportunities to increase their vocabulary proficiency before their departure and if post-tests are delayed for too long after the return then the chance of vocabulary knowledge attrition would increase. It was noted that in Fitzpatrick's (2003) study her subjects received their initial pre test only after they had spent the first weekend with their host family. During this two-day period, although short, there was some

opportunity for latent vocabulary to be reactivated. As noted previously in chapter 2, Meara (2005) discussed evidence for the spontaneous reactivation of vocabulary knowledge and looked at data suggesting that his test subject's active vocabulary more than tripled in size over the course of just two days. His results should be treated with caution as he warned that they were not conclusive but it does seem that exposure to a L2 environment can quickly to encourage the reactivation of more frequently occurring words in particular. He also found that high frequency words were more likely to be encountered at first upon initial exposure than low frequency. Although Meara failed to find conclusive evidence he suggested "that we need to be aware that rapid and extensive vocabulary reactivation as a result of environmental input needs to be taken more seriously" (2005, p. 279).

3.5.5 Scoring protocols

I referred briefly to the lemmatization procedure earlier in the paper where we noted criteria used by Meara and Fitzpatrick (2000) which corresponded to levels 2 and 3 of Bauer and Nation's "word families" (Bauer & Nation, 1993). I tried to award a mark for each infrequent vocabulary item and give "credit at every possible opportunity" (Meara & Fitzpatrick, 2000, p. 26) but processing each test subject's word corpus was sometimes problematical. Clenton (2005, p. 53) gives some examples from a Japanese context including the use of *katakana* (Japanese syllabic writing primarily used for words of foreign origin) in some of the responses and the use of loan words which, in practice, are used very differently in Japanese. Jiménez Catalán and Moreno Espinosa (2005) also looked at similar issues with Spanish students. I tried to overcome some of these difficulties by compiling a list of protocols (see Appendix 1) and attempting to be as consistent as possible with their application.

A further point can be made about test responses. Fitzpatrick (2003) discusses the Edinburgh Associative Thesaurus (Kiss et al., 1973) at some length explaining how the 30 cue words for the Lex30 test were selected from more than 8,000 possible options. Given the protocol used by the Edinburgh Associative Thesaurus for cue word selection in which Native Speaker (NS) associative responses were used it seems likely that some degree of bias will exist. The degree to which Japanese L1 English learners are more likely to produce different, but many might

argue valid, responses than NSs (Dewey, 2012) will be discussed in more detail later in this thesis.

3.5.6 Percentage score v. Raw score

The last issue to be considered is the method of scoring. In earlier pilot testing Fitzpatrick (2003) and in the Fitzpatrick and Meara's (2004) study the raw Lex30 score was used as a basis to measure performance. She soon noticed, however, that there was a much greater variation in the number of words produced in response to Lex30 than there had been with earlier versions of the test and she became concerned that the Lex30 was tending to emphasize corpus size over quality (in terms of high-frequency words). She felt that any performance score calculated from the test task should be as independent as possible and should not allow the number of words produced to affect measurements of the quality of words. As a result the raw Lex30 score was recalculated in terms of the number of infrequent words as a percentage of the total number of words produced therefore reflecting the proportion of infrequent words in each corpus. In both Fitzpatrick's (2003) study and in the replication there was a tendency for participants to produce far more words at test Time 2 than at test Time 1 which might be, as expected, associated with language study between the two test times during both longitudinal experiments. In terms of the mean total number of words produced, Fitzpatrick (2003) study group increased from 73 to 94, a rise of 21 while the replication study group increased from 48 to 74, a rise of 26. This suggests that there was rise in vocabulary production (which might be a proxy for fluency) within both groups. The mean number of infrequent words produced by all subjects also increased from 17 at test Time 1 to 20 at test Time 2 (Fitzpatrick, 2003) and from 18 to 26 (replication). In both studies, figures for both the total number of words produced and the raw Lex30 scores increased but because the total number of words increased by considerably more the percentage score fell in both cases. This indicates that the increase in learners' production of frequent words might be because those words were shifting from receptive to productive knowledge or had acquired more associations or connotations or collocations as a result of SA - in other words, that SA generates greater depth of knowledge of words already acquired, as well as helping with the acquisition of new items.

3.6 Conclusion

The replication experiment tried to reproduce the results of two previous longitudinal experiments with mixed success. Comparisons with Clenton and Fitzpatrick (2010) were encouraging. Using a similar number of participants, similar test administration and protocols and most importantly the same Lex30 scoring system using a raw count of infrequently occurring words, the results that I obtained were roughly equivalent. With Fitzpatrick's (2003) study I was forced to reconsider what exactly I was measuring and to reexamine the data that I had gathered. Is having a raw Lex30 score sufficient to make a judgment of subject's improvement in productive vocabulary knowledge over the course of a short-term SA programme or should account be taken of the influence of the total size of the corpus produced? Perhaps further exploration into how data from the Lex30 is processed will help us formulate new and better-balanced marking schemes for the future.

Chapter 4: A closer look at Lex30 scores before and after SA

In this chapter I will describe an experiment that uses Lex30 to collect data on Study Abroad (SA) participants' productive vocabulary knowledge. Examining such data might help us reveal if any changes take place during an SA experience. The replication study in chapter 3 goes some way towards suggesting that SA participants can develop an ability to generate a higher number of low frequency words within a relatively short period of time.

In the previous chapter I reported a replication study which, like previous studies (e.g., Fitzpatrick & Meara, 2004; Fitzpatrick & Clenton, 2010), identified changes in learners' productive vocabulary performance after a study abroad period. However, questions were raised about the nature of these changes, and the ways in which Lex30 scores might best be calculated in order to capture information about vocabulary development. The study reported in this chapter offers more detailed scrutiny of the changes in Lex30 performance. It also includes two extra test events, approximately two/three weeks before departure and approximately two/three weeks after return from study abroad. This helps us to gain a longer view of vocabulary development, and to understand the extent to which changes in vocabulary knowledge as a result of SA are retained.

The next part of the chapter will describe the methodology used, particularly with reference to scoring protocol and task administration, and make comparisons with that used in past applications of this measurement tool. In addition to using previously used methods to analyze the results I will also suggest some new ways of enhancing our understanding of them. For instance, as well as considering two simple categories of either frequently occurring or non-frequently occurring word types, I will also examine the use of multiple narrow word frequency bands as proposed by Kremmel (2016) to see if fine-grained frequency analysis can reveal more about vocabulary knowledge development.

Finally, during the discussion, I will try to account for some of the changes in the number of words produced by participants at various time points during their SA experience and attempt to explain why different methods of scoring Lex30 can yield such different results. I shall also suggest improvements that might be made in both the design and application of a language measuring tool like Lex30 based on the results obtained.

4.1 Lex30 administration

In previous experiments (Fitzpatrick, 2003; Fitzpatrick & Meara, 2005; Fitzpatrick & Clenton, 2010) the Lex30 task has been given on two occasions immediately before and after an intervention event such as an intensive language learning course. In the experiment reported in this chapter, data is gathered at multiple time points - in this case a total of four will be used. The first occurs approximately three weeks before departure (early pre-departure test), the second on departure (pre-departure test), the third immediately on return (post-test) and the final one approximately three weeks after return (delayed post-test).

In looking at the difference in data gathered at post-test and delayed post-test time points the study breaks new ground. A review of research has so far found no instance where a delayed post-test has been used with Lex30. A delayed post-test is likely to reveal if any decline or attrition in vocabulary takes place after a learning event such as in our case, an SA experience. Previous studies have looked at learners' vocabulary knowledge attrition using a number of alternative methods of measurement. Ecke and Hall (2012) conduct a case study into rates of vocabulary attrition of a multilingual speaker. They show that even a speaker's L1 can undergo mild attrition when competing with more recently learned languages. Schmitt (2010) distinguishes between the attrition of receptive and productive mastery of lexical terms and presents evidence to support the view that productive knowledge tends to attrite more than receptive knowledge particularly when it comes to low frequency words. The rate of attrition of vocabulary knowledge might also be connected with proficiency level. Findings from research by Hansen et al. (2002) indicate that learners with larger vocabularies might retain their knowledge more effectively over time. There is evidence to suggest that the process of language attrition does not continue indefinitely and that some aspects of language knowledge could be retained by some learners over very long periods (Weltens et al., 1989). Given this evidence it seems likely that SA learners, on return to their home L1 environment, would be likely to experience a similar loss of vocabulary. Particularly in the case of low proficiency participants this might mean losing their mastery of low frequency L2 words.

4.1.1 Scoring, word-bands and data analysis.

With the use of Lex30 in the experiment described in this chapter, I plan to follow the same protocol as seen in most previous uses with regard to scoring. That is to say, a point score is given for every response which is classed as an infrequent word, with the definition of “infrequent” being outside the first 1000 most frequent English words. For more details of scoring protocol including previous examples used by a number of researchers please refer to sections 2.8.1, 3.1 and Appendix 1. As I have noted, unlike other frequency-based tests, Lex30 does not make any distinction between bands other than the first 1000-word band marker where scoring protocol divides words produced by test-takers into just two groups. Although adopting this simple form of classification may aid the calculation and comparison of scores, there is also a risk that this oversimplification of word-bands and the use of a single 1K/1K+ threshold might lead to missed opportunities when looking at smaller changes in vocabulary knowledge.

Kremmel (2016) uses a range of different width bands to classify word frequency arguing that bands differ in their importance in terms of the coverage they provide. Since a large proportion of the words that learners produce are likely to be more frequently occurring he suggests that smaller 500-item bands might be more informative at higher frequencies, and bands larger than 1,000 items would be adequate at lower frequencies.

4.2 Research Questions

In order to investigate whether Lex30 can detect changes in vocabulary knowledge as the result of an SA experience this study sets out to address the following three research questions:

1. Is there any change in the total number of words SA participants produce before, during and after an SA experience?
2. a) Is there any change in Lex30 score produced by the same participant?

b) Is the change in test performance over the SA period (at test Times 2 and 3) significantly different from the change in test performance at test Times 1 and 2?

- c) Do learners perform similarly at test Times 3 and 4, or is there evidence of attrition?
3. Is a more fine-grained application of word frequency analysis (Kremmel 2016) more informative and can it contribute further to an investigation of vocabulary change during SA?

4.3 Methodology

4.3.1 Participants

41 L1 Japanese female EFL learners participated in a 15-day SA programme in London in the UK. All were first year students studying at Nakamura Gakuen University in Fukuoka, Japan and were aged between 18 and 19 years old. Most of the group had no internationally recognized English language proficiency qualification. The average learner was estimated to be around advanced beginner level or CEFR (Common European Framework of Reference for Languages) Upper A1. The CEFR is widely accepted as being the global standard for grading an individual's language proficiency and the Upper A1 level is seen as roughly equivalent of a TOEIC score of between 230 – 280. In common with many Japanese university students, all group members had studied English as a compulsory subject at school for at least six years.

The learners were divided into two groups depending on area of study. The first comprised of 25 students belonging to the nutrition department of the University studying to gain qualifications as dieticians or in food sanitation management. The second group of 16 students belonged to the careers development department which specializes in bookkeeping, computing and other business management skills. Both groups receive a similar number of three 90-minutes English classes per week.

4.3.2 SA programme

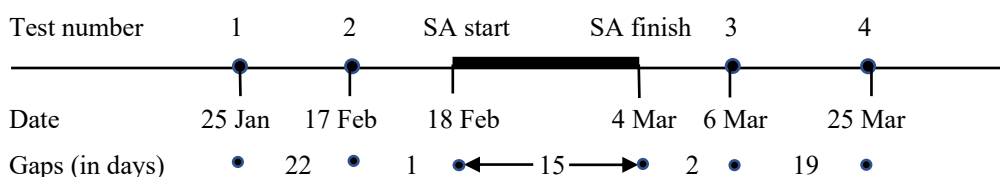
The SA programme lasted 15 days and took place at the Eurocentres School in Victoria, London between 18th February and 4th March 2017. The participants attended three 50-minute morning custom-made English classes for a total of nine days. The group was divided into three separate classes depending on ability level and decided by a simple level check test given on the first day of attendance.

Eurocentres also arranged an opportunity to do a one-day work internship with the charity organization Oxfam. During the weekends and the afternoons of most weekdays the participants took part in a variety of activities ranging from a London open top bus tour to lessons in flower arrangement and afternoon tea all supervised and taught by local English L1 speaker specialists. A day-out was spent visiting Cotswolds and Oxford on a private bus tour.

The overriding emphasis of the SA programme was to maximize the amount of out-of-class contact opportunity where participants can practice their English in a natural setting. An important part of this is having students stay with individual host families where they are likely to have opportunities to socialize. Nakamura Gakuen University lays down a strict policy that no two Japanese students should stay with the same family as this is thought to discourage English communication. The Eurocentres school also provided a great deal of helpful encouragement to students preparing them for out-of-class contact opportunities by teaching and practicing a range of functional language items designed to be used in a variety of practical situations. Past research has underlined the importance of preparing students in this way. Briggs (2015) found that there is a connection between out-of-class language contact and vocabulary gain in SA contexts. She advises that language institutions hosting SA learners can do much to guide the out-of-class contact of their learners and that this can have clear pedagogical implications. Kinginger (2011, p. 67) also highlights the important role the institution can play in promoting the engagement of SA learners in out-of-class L2 contact by encouraging them to take linguistic advantage of that contact. Finally, Baker-Smemoe et al. (2014) provides evidence to show how schools can adapt an overall curriculum in ways which can further encourage out-of-class L2 contact and engagement.

4.3.3 Test administration

A decision was made to use the written version of the Lex30 in much the same way as has been used in previous research. A computer-based version where students input their responses to cue words using a personal computer was briefly considered but then deemed impractical. With travel outside Japan there was no guarantee that computer facilities would be available and most participants did not have access to their own personal computer or tablet.

Table 4.1*Test administration and number of days between tests*

A standard version pen-and-paper written version of the Lex30 task, which has been previously described in section 3.3 and by Fitzpatrick and Clenton (2010), was administered on four occasions as shown in Table 4.1. The gaps between test times varied between 18 days (test Times 2 and 3) and 22 days (test Times 1 and 2) as there were difficulties in arranging suitable times for test-takers to meet. It was thought important for everyone to complete the test at the same location under identical conditions to ensure as much consistency as possible.

4.3.4 Lex30 cue words and scoring protocol

The fundamental design of Lex30 has remained constant (see Figure 2.7, p. 50) although some of the ways in which this measurement tool has been scored have varied. In early examples (e.g., Meara & Fitzpatrick, 2000; Fitzpatrick, 2003; Fitzpatrick & Meara, 2004) Nation's (1984) word lists were used both for selecting cue words and scoring their responses. More recently the JACET 8000-word frequency list has been used (JACET, 2003) for scoring purposes (e.g., Jiménez Catalán & Moreno Espinosa, 2005; Fitzpatrick & Clenton, 2010; Clenton, 2010; Fitzpatrick & Clenton, 2017; Walters, 2012). In the experiment described in this chapter and in the replication experiment described in chapter 3 (see section 3.3) the same JACET 8000 word frequency list has been used. For Japanese test-takers, who are perhaps more used to modern American-English vocabulary examples, this seems an appropriate policy.

The protocol used with scoring in this experiment again closely follows previous applications of the Lex30 task. Test-takers produce up to four responses to each of the 30 cue words although in many cases this maximum is unlikely to be reached. Each of the responses is lemmatised so that suffixes such as plural forms, past tenses or comparatives and frequent regular affixes (-able, -ly, etc.) are counted

as examples of base-forms of these words. The guidelines used by Meara and Fitzpatrick (2000, pp. 29-30) were followed with one or two specific exceptions.

Every response was carefully examined using the JACET first 1000-word list. Every word which occurs on this list (<1K) was awarded '0' points and answers which do not occur on the list (infrequently occurring or 1K+ words) but were deemed 'acceptable' are awarded '1' point. Acceptable words do not include items such as proper names, acronyms and non-English words. A detailed description of both acceptable and unacceptable words is given below. Generally, with one exception, the same protocol is used as in the replication experiment (see Appendix 1). The exception is with point (7) on the 'acceptable' word list which includes countries, languages and nationalities. Words in this particular category are considered <1K words and are awarded '0' points. They are also included in the count of 'total number of responses' given.

In producing their answers it seems natural that test-takers are likely to make spelling errors in producing some of their responses. In these cases the overall general policy is to give as much credit as possible for whatever attempt was made as long as the word is recognizable. For the purposes of this experiment on occasions when word recognition might be called into question, an arrangement was made where three English L1 speakers would examine each word and, if unanimous agreement is reached regarding the intended word, the word is accepted. The number of misspelled (but acceptable) words and discounted words produced during each test event is also noted.

Acceptable words

Words are deemed acceptable and are included in either the 1K or 1K+ word groups in the following cases:

1. Misspelled but nevertheless recognizable words as discussed above. Jiménez Catalán and Moreno Espinosa (2005) argue against this saying that there should be greater score weighting for correctly spelled answers. However, as I am interested in the threshold knowledge of individual words in addition to more comprehensive vocabulary knowledge, I decided to allow word inclusion wherever possible even in cases where correct spelling may not yet be fully acquired.

2. In some cases identical words were produced in response to two or more cue words. An example might include the word *delicious* in response to the cue words *potato* and *rice*. This was judged as acceptable as long as some kind of semantic relationship exists between the response and the cue.
3. In other cases test-takers sometimes used the cue provided as responses to some of the other cues (Jiménez Catalán & Moreno Espinosa, 2005, p. 41). For this experiment responses given in this way were only deemed acceptable if there is again some sense of a semantic relationship. For example the cues *pot* or *potato* given in answer to the cue word *rice* is permissible but not *attack* or *furniture* as in this case any semantic link is questionable.
4. Where a two part ‘phrasal verb’ or other multi-word expression is written in a space meant for a single word answer a score will still be given. Particularly in the case of pen-and paper version of Lex30, findings from previous experiments tell us that test-takers are likely to do this on occasion in spite of clear test instructions⁴. For example, in response to the cue word *hold* the expression *take place* would be score ‘0’ as each of the constituent words ‘*take*’ and ‘*place*’ appear on the JACET 1000-wordlist. However, the expression *take a shower* written in response to the cue word *habit* scores ‘1’ point as ‘*shower*’ does not appear on the same wordlist. Likewise with the expressions *staple food* and *good taste* given in response to the cue word *rice*. Both word pairs score ‘1’ as ‘*staple*’ and ‘*taste*’ occur outside the first 1000-words. When the two or more words written in the same space are 1K+ words such as *dietary fiber* (given in response to the cue word *potato*) they will still be only counted as a single word and are credited with ‘1’ point.
5. Words written using either American or British English spellings.

⁴ This complication is less likely to occur when the task is administered by computer since there are automatic constraints on the form of responses test-takers are allowed to give.

6. Words which appear on the “List of English words of Japanese origin” from Wikipedia (as of 1st January, 2020) are considered acceptable. For example the words *sumo*, *judo* and *sushi* are counted but not *kanji* or *hiragana*.

7. Countries, languages and nationalities like *Japan*, *Japanese*, *England* and *English* score ‘0’ points. This differs from Jiménez Catalán and Moreno Espinosa (2005) who argue that proper names of countries, when written in English by low level learners, indicate some level of L2 productive knowledge and accepted words in this category with a score of ‘1’ point. My reason for not giving greater credit in the present experiment is that words of this type can be produced as a primary response to certain cue words more often than would be expected under normal (or non-SA) conditions. This could be explained by the fact that the cultural background of the test-taker along with their SA experience encourages them to produce words of this type much more frequently than they would otherwise do so. For example, during SA, learners will be more likely to make comparisons between their own culture and that of their new L2 environment prompting them to identify with and label certain nouns in a certain way. For example *English* was written down a total of 41 times in response to the cue *spell*, *Japan* or *Japanese* was given 38 times in response to the cue *rice* and so on. As countries and country adjectivities do not appear on the JACET list of 1000 most frequently occurring words, they thereby attract a higher score. This means that, for many test-takers, there is a risk of gaining a high overall Lex30 raw score which might result in overestimating the real level of their productive vocabulary knowledge. For this reason it was decided not to award a ‘1’ score with these particular word types but still include them in the ‘total number of responses given’.

Unacceptable words

In the following cases words produced by test-takers are non-scoring words and are not considered acceptable. These include:

1. Proper nouns like *McDonalds*, *Victoria* or *Kentucky*.

2. Acronyms such as DVD, CM and PC.
3. Numbers are also not counted either written numerically (1, 50, 100) or in words (one, two, three). The reason for this was to avoid a tendency for test-takers to merely compile ‘number lists’ instead of demonstrating their knowledge of a wider range of different lexical items.
4. Finally, Japanese words which do not appear on the “List of English words of Japanese origin” from Wikipedia.

4.3.5 Methods of analysis

In order to address the research questions set out in section 4.2, I undertook the following analyses, the results of which are presented in the following section:

1. I calculated the number of total number of responses given (with both acceptable and discounted words) at each of four time points. Then I examined the number of misspelled (but acceptable) words and their proportion in relation to the total number of words produced. This shows if there is any tendency toward improvement in spelling accuracy in absolute terms. This topic will be more closely examined in chapter 6.
2. I calculated descriptive statistics gathered from Lex30’s results at four time points including the mean of total words and infrequently occurring words (i.e. the Lex30 raw scores) produced along with their standard deviations. The data was analyzed using one-way repeated measures analysis of variance (ANOVA). This was used to determine whether there are any statistically significant differences between the means of three or more levels of a within-subjects factor. Repeated paired t-tests analyses can lead to high false discovery rates over time. Since our experiment involves data gathered over four test times use of repeated measures ANOVA with Bonferroni correction takes into account the likelihood of a high false discovery rate when a number of simultaneous statistical tests are done.
3. Finally, I re-categorized the words produced by test-takers at all four time points into both narrower (high frequency) and wider (low frequency) word bands as

described by Kremmel (2016). This might reveal more detailed shifts in productive vocabulary knowledge during an SA experience.

4.4 Results

4.4.1 General overview

A total of 41 Japanese L1 EFL learners took part in the SA programme. Due to personal reasons three of them were unable to complete all four of the Lex30 tasks. Rather than trying to accommodate incomplete data sets, a decision was taken to remove all their data from analysis, leaving us with 38 participants contributing a total of 9437 words for our analysis. I have included a completed Lex30 task paper from an individual participant (participant 22) in Appendix 2. The participant completed the task at Time three, 2 days after their SA experience.

Table 4.2 shows the totals of words produced at each time point on the left and the number discounted according to the criteria in section 4.3.4 above before leaving a net total. The number of words belonging to the mostly frequently occurring (<1K) and less frequently occurring (1K+) word bands are shown on the right of the table. There is a steady increase in “score” across Times 1, 2 and 3, and then a small decline at Time 4. And for all of them, the difference between Times 2 and 3 is the most marked.

Table 4.2

Total words produced by participants during their SA experience

Time point	Total words	Discounted	Net total	Mis-spelled	% total	<1K	1K+
1	1833	102	1731	211	12.19	1018	713
2	2094	112	1982	200	10.09	1187	795
3	2841	136	2705	216	7.99	1611	1094
4	2679	90	2589	219	8.46	1580	1009
Total	9437	440	8997	846		5296	3611

In these results I will firstly present the overall results of the experiment showing the number of frequent and infrequent words produced by 38 participants over 4 time points as well as providing details of discounted and misspelled words. Secondly, I will investigate if there were any significant changes in the total number

of words and in the number of infrequent words (Lex30 raw score) produced at each of these time points. Finally, I will examine if a more fine-grained application of frequency analysis can reveal further changes in vocabulary knowledge development.

Figure 4.1

Total words produced at each time point by all 38 SA participants.

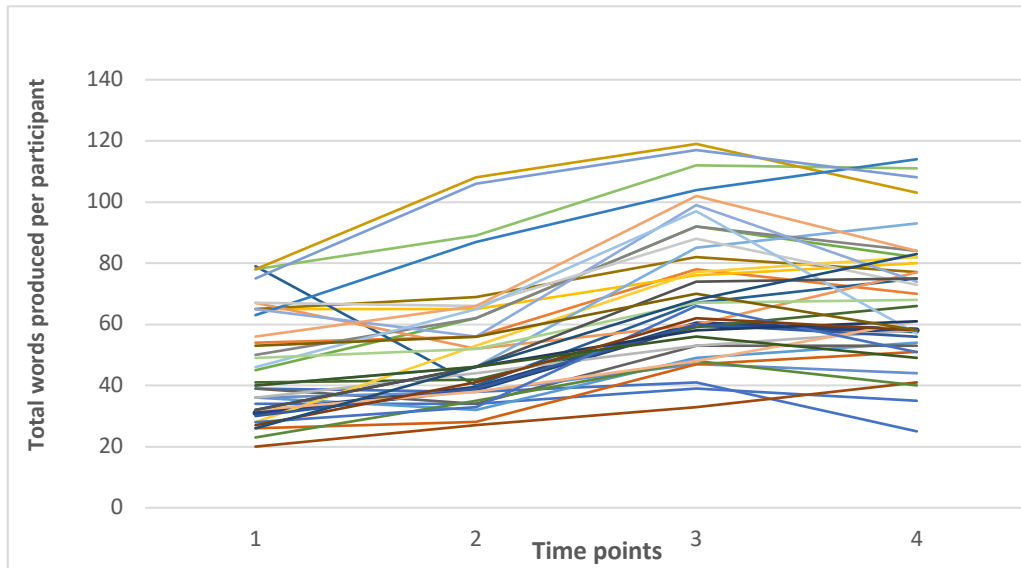
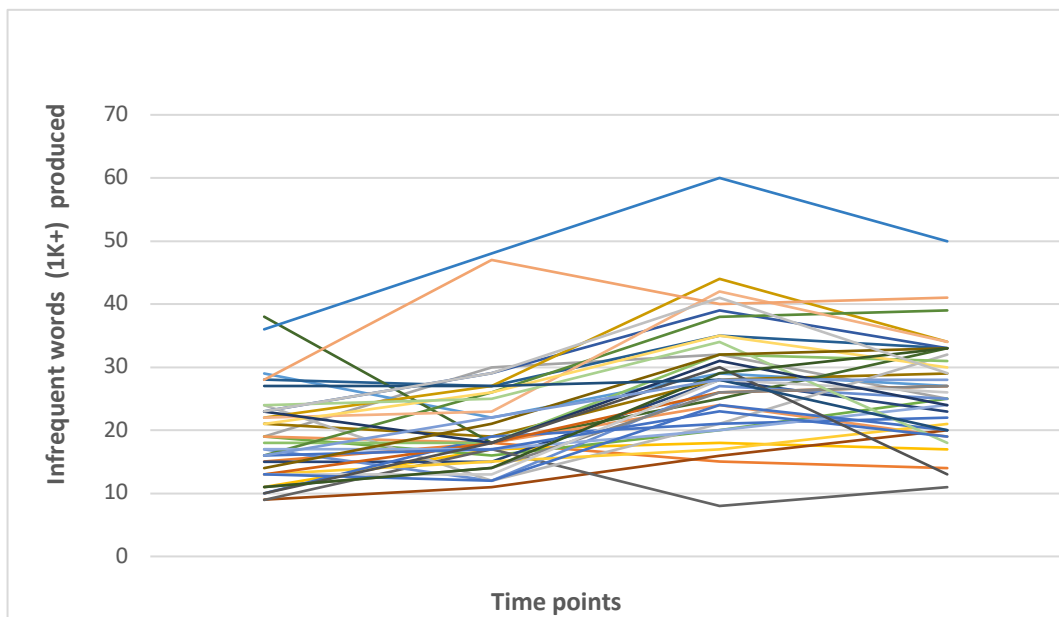


Figure 4.2

Total infrequent words (1K+) produced at each time point by all 38 SA participants



Moving on to look at the results at the participant (rather than group) level, Figure 4.1 shows the total number of words produced by each participant at each of the four time points. Each of the 38 participants is represented by a different coloured line on the graph. The lines indicate that most participants mirror the trend in the group data, with a rise in the number of words produced between Time points 1 and 2, a rather greater increase between Time points 2 and 3 and finally, a fall in the total number of words produced between Time 3 and 4. Figure 4.2 is a similar graph showing the number of infrequent (1K+) words produced by each participant between Time points 1 and 2, a rather greater increase between Time points 2 and 3 and finally, a fall in the approximately mirroring the same trend. There is some evidence of extreme data points or ‘outliers’ on the second graph particularly with the highest and lowest scoring participants.

Table 4.2 shows that SA participants produced a total of 9437 words on the four occasions the Lex30 task was administered. Of these 440 words were discounted for the reasons given below. Around 70% out of the total 440 words were discounted due to unrecognizable spelling. Three English L1 speakers involved in full time education in Japan examined the misspelled words test-takers had produced. In each case, where no unanimous decision could be reached, the word was rejected. For example, in response to the cue word: **attack**, responses like *spote*, *boal* and *vocucing* were rejected while *vollyball*, *quarel* and *ragby* were allowed. In addition, for the cue **potato**, *crocat*, *jank* and *sitew* were not recognized while *deliisious*, *begetable* and *fride* were included.

Again, looking at Table 4.2, it is interesting to note that there is a gradual decrease in the percentage of misspelled words compared to the total number of words produced during the test-takers’ SA experience. At Time 1 the percentage is 12.19% decreasing to 10.09%, at Time 2 and 7.99% at Time 3. After the return to Japan, at Time 4, this number slowly rose once again to 8.46%. More details of orthographical changes during SA and their possible implications can be found in chapter 6.

Around 15% of words included in the 1K group (and awarded ‘0’ points) were place and country names, adjectives of nationality and languages. For example *English* was written down a total of 41 times in response to the cue *spell*, *Japan* or *Japanese* was given 38 times in response to the cue *rice*. For the reason explained earlier (see section 4.3.4) words of this type not given credit as infrequent (+1K)

words as they were produced more often than would normally be the case in response to certain cues. To avoid any risk of overestimation of test-takers' overall Lex30 raw scores it was decided to award '0' instead of '1' point for these particular word types.

Around 10% of words in the discounted word group were made up of proper names. Examples like *Google* and *Yahoo* were produced numerous times in response to *map* and *MacDonalds* to the cue word *potato*. In about 5% of cases where words were discounted, cue words from the 30 provided were given in response to different cues. Where some semantic relationship seems to exist this was deemed acceptable. Examples include cue word pairs like *close* or *window* and *rest* and *seat* where cue words were given in response to each other. Another example was the cue word *pot* written in response to *rice*. Sometimes, however, inappropriate cue words were used. Most examples occurred when the same cue word was written down as a response to its original. The remaining 5% of discounted words were made up of acronyms. *TV*, *DVD*, *CM* and *SF* were commonly written down in response to the cue word *television* and *H2O*, *DNA*, *CO2* and *O2* in reply to *science*. In addition individual letters such as *a*, *b*, *c* and *d* were written for *spell*.

Two more points can be made about the responses given in the Lex30 task. The first is that sometimes more than one word was written in a single space which meant a decision had to be made which of the extra word(s) should be discounted. The process for dealing with this occurrence has already been outlined (see section 4.3.4). It was found certain cue words attracted more "multiple word single space" responses than others. For example the cue *habit* attracted responses such as 'take a shower' 'watch a movie' and 'read a book' and the cue *rice* word combinations like 'curry rice' or 'rice bowl.'

The second point, previously referred to in section 4.3.1, is that infrequently occurring scientific words were produced by some of the test-takers particularly from those studying nutrition as a major part of their university studies. Examples included *nutrition*, *protein*, *dietician* and *biochemistry*. However, while interesting, the number of such words was small and not thought to affect the overall result.

4.4.2 Changes in the total number of words produced (RQ1)

In answer to the first research question, the total number of words produced (8997) for the Lex30 task over 4 time points, was analysed. As seen in Table 4.3 below, the mean number of words produced by SA participants (N=38) gradually

increased from 45.55 words to 52.16 words (Time 1 to Time 2) to 71.18 words (time 3) before decreasing slightly down to 68.13 words (Time 4).

Table 4.3

Descriptive analysis: total words

	N	Min	Max	Mean	Std deviation
Time 1 total words	38	20	79	45.55	17.387
Time 2 total words	38	27	108	52.16	19.692
Time 3 total words	38	33	119	71.18	22.646
Time 4 total words	38	25	114	68.13	20.903

A one-way repeated measures ANOVA analysis was conducted for research questions one (1) and two (2) to determine whether there were statistically significant changes in the total number of words and the number of infrequent words produced by participants before, during and after a short-term SA programme. The data for both experiments was examined before the final analysis was conducted. Although measurement variables were not normally distributed in both cases, this may increase the likelihood of a false positive result if when analyzed with a test like repeated measures ANOVA that assumes normality. Fortunately, ANOVA is not very sensitive to moderate deviations from normality. Research with simulation studies, using a variety of non-normal distributions, have shown that the false positive rate is not affected very much by this violation of the assumption (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). Finally, Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated for both analyses.

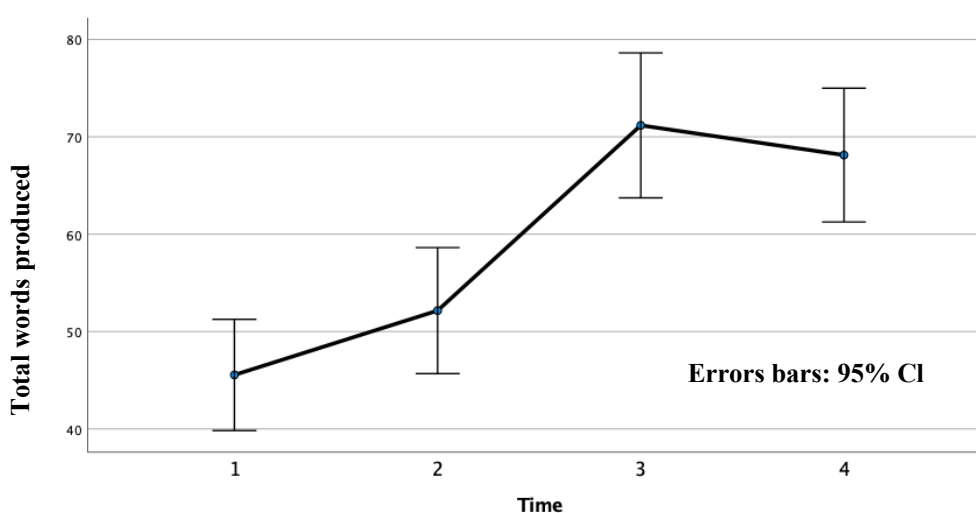
For research question one (1) a one-way repeated measures ANOVA was conducted to determine whether any differences in the total number of words produced at four time points over the course of an SA programme were statistically significant (see Appendix 3 for the SPSS results). There was a statistically significant effect of test time on the number of words produced [$F(3, 111) = 66.624, p < .001$]. The effect size, calculated as eta squared (η^2), was 0.643, indicating a large effect. Post hoc analysis with a Bonferroni adjustment revealed that an increase in the total number of words produced by participants was statistically significant from Time 1 to Time 2 (-6.61 (95% CI - 12.30 to - .96) $p = .015$) and from Time 2 to Time 3 (-

19.03 (95% CI - 23.47 to - 14.59) $p < .001$. However, no significant difference was found between Time 3 and Time 4 (3.05 (- 2.18 to 8.29) $p = .674$).

These results can be graphically demonstrated when looking at Figure 4.3. We can see that the 95% Confidence Intervals (CI) for Time 1 and Time 2 and for Times 3 and 4 both seem to overlap. However, the intervals for Time 2 and Time 3 do not which indicates the size of the difference in number of words produced by learners between those test times. In answer to the first research question the findings indicate that there is change in the total number of words produced before, during and after an SA experience with significant increases taking place between Times 1 and 2 and between Time 2 and Time 3.

Figure 4.3

Lex30 task: total words produced over time



4.4.3 Changes in the number of infrequent words (RQ2)

Table 4.4

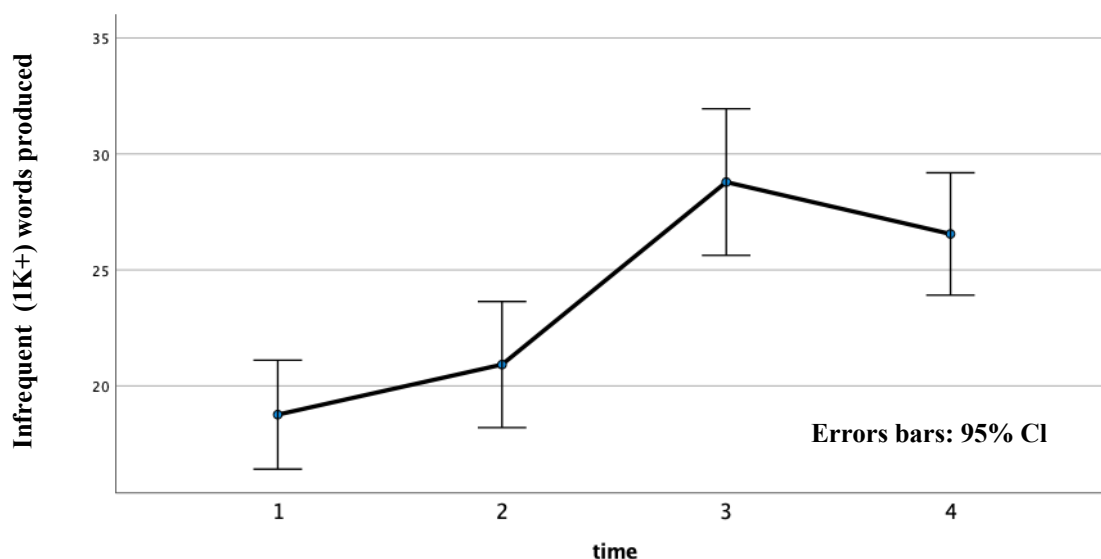
Descriptive analysis: infrequent words

Time point	N	Min	Max	Mean	Std deviation
Time 1 Infrequent words	38	9	38	18.76	7.145
Time 2 Infrequent words	38	11	48	20.92	8.270
Time 3 Infrequent words	38	8	60	28.79	9.612
Time 4 Infrequent words	38	11	50	26.55	8.026

In relation to the second set of three research questions, data which looks at changes in the number of infrequent words produced in the Lex30 tasks (i.e. the Lex30 raw scores) over four time points, was analysed. As seen in Table 4.4 below, the mean number of infrequent (1K+) words produced by SA participants (N=38) gradually increased from 18.76 words to 20.16 words (Time 1 to Time 2) to 28.79 words (Time 3) before decreasing to 26.55 words (time 4). A one-way repeated measures ANOVA was conducted to determine whether any differences in the total number of words produced at four time points over the course of an SA programme were statistically significant (see Appendix 4 for SPSS results).

Figure 4.4

Lex30 task: infrequent (1K+) words produced over time



There was a statistically significant effect of test time on the number of words produced [$F(3, 111) = 34.870, p < .001$]. The effect size, calculated as eta squared (η^2), was 0.485, indicating a large effect. Post hoc analysis with a Bonferroni adjustment revealed that a change in the total number of words produced by participants was not statistically significant from Time 1 to Time 2 (-2.16 (95% CI - 5.24 to .928) $p = .353$) and from Time 3 to Time 4 (2.24 (95% CI - .52 to 5.00) $p = .179$). However, a significant difference was found between Time 2 and Time 3 (-7.87 (95% CI - 10.70 to - 5.04) $p < .001$). These results can be graphically demonstrated when looking at Figure 4.4. We can see that the 95% confidence

intervals for Time 1 and Time 2 and for Times 3 and 4 both seem to overlap in a similar way to Figure 4.3. The confidence intervals for Time 2 and Time 3 do not overlap which is an indication of the size of the difference in number of infrequent words produced by learners between those test times. In answer to the second research question: (a) there seem to be changes in the Lex30 score produced by the same participant, (b) the change in test performance over the SA period (at test Times 2 and 3) is significantly different from the change in test performance at test Times 1 and 2 and finally, (c) there does seem to be evidence of some attrition in Lex30 performance between test Times 3 and 4 although this is not significant.

4.4.4 Fine-grained word frequency analysis (Kremmel 2016) (RQ3)

The previous section focussed on the number of infrequent words produced at each test time, with infrequent defined as “not within the first thousand band”. In order to address this third research question, I took a more fine-grained approach to frequency, using 500-word bands for the first 3000 words, and 1000-word bands until the 6000-word mark and 2000-word bands thereafter (up to 8000). As discussed in section 4.1.2 above, these distinctions are informed by Kremmel (2016).

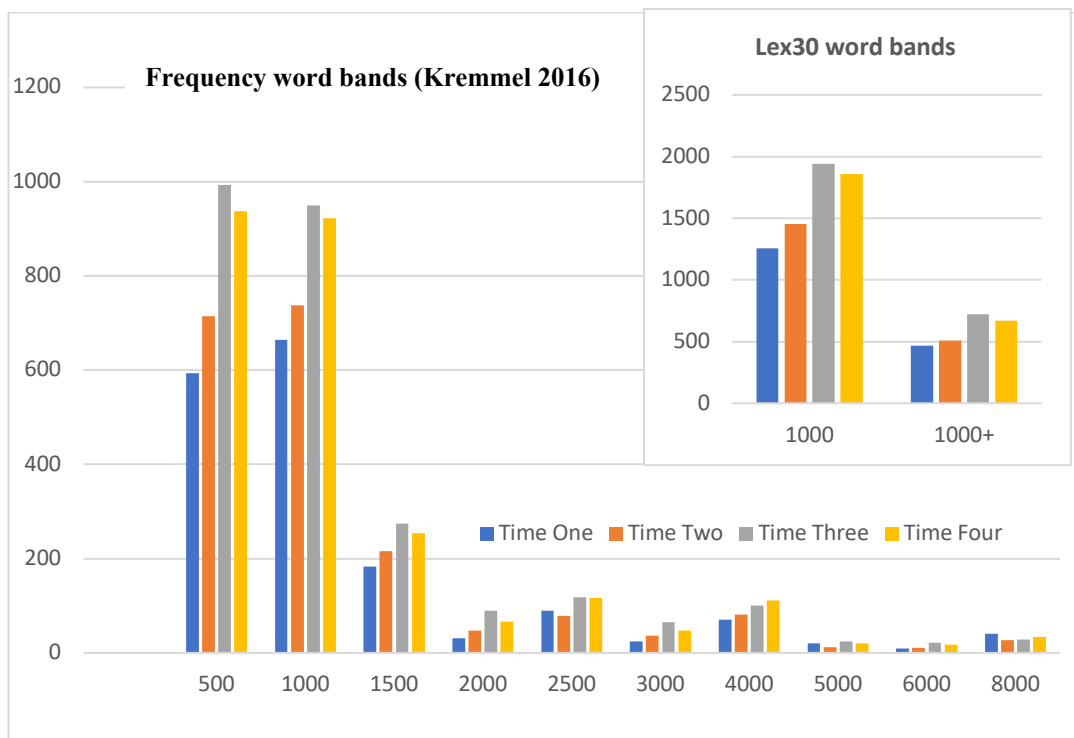
Words gathered using the Lex30 task were divided into word bands according to frequency. Figure 4.5 (p. 93) shows the total number of words produced by 38 SA participants at all four time points. The superimposed graph in the top right corner shows the words divided into just two word bands, 1K and 1K+, which follows the usual scoring system adopted by the Lex30 task. On the smaller graph, the number of high frequency words produced at each time point increases from Time point 1 to Time point 2 with a much greater increase between points 2 and 3. Finally, there is a slight decrease from Time point 3 to 4. The same pattern is observed with less frequently occurring words (1K+) although the total number of words involved is considerably less. The main graph shows the same data divided into a greater number of word bands as proposed by Kremmel (2016). Looking at the most 1000 most frequently occurring words we can see that there is little difference between the first 500 and the second 500 word bands although there is perhaps a slightly higher number of words produced in the first band. There is a substantial drop in the number of words produced between 1000 and 1500 word band points. The graph still traces the familiar pattern of a small following by larger increase up until Time 3 followed by a gradual decline towards Time 4. The same occurs again

with the next frequency band (1500 to 2000 words). However, from the 2500 word band onwards, the pattern appears more random all the way to the 8000 word band.

In answer to the third research question it seems that the fine-grained application of word frequency analysis has the potential to be more informative than Lex30's simple 1K/1K+ word band division but in the case of our experiment it seems to offer little in the way of new knowledge. Perhaps using a larger data base with words representing a wider range of proficiency levels this method might be more likely to reveal information about vocabulary development during SA.

Figure 4.5

Comparison of fine-grained word bands with Lex30 word band divisions



4.5 Discussion

In this section I will firstly try to account for the changes seen both in terms of total words and infrequent words, produced at four time points before, during and after an SA experience. Secondly, I will revisit Kremmel's (2016) detailed word frequency bands and see if they can help distinguish finer and more meaningful changes in the frequency of words that SA participants are likely to produce. Finally, I will briefly outline some changes that might be made with both the design and application of Lex30 based on the results obtained.

4.5.1 Changes in the number of words produced

With both the total number of words and the number of infrequent words (1K+) produced, there appears to be a similar pattern, showing that these measures are not independent of each other. If participants produce more words, it is then likely that some of those will be infrequent words. Looking at Figures 4.1 and 4.2 (p. 86) there is an increase between Time 1 and Time 2 followed by a much larger increase at Time 3. Then there is a decline from Time 3 to Time 4. A possible explanation exists for why these changes might occur in each case.

Firstly, in the few weeks before their departure for the UK, the participants took part in series of preparation activities including English language classes which were given at least three times over the course of a 6-week period. Added to this there were likely to be some feelings of anticipation before undergoing an SA experience. Some students admitted to studying English privately as they were nervous about conducting conversations with their respective host families. Such preparatory behaviour on the part of the SA participants was unanticipated and might go some way towards explaining the apparent rise in the number of both total and infrequent words they were able to produce.

Secondly, my analysis found that there was a significant difference in the total number of words and infrequent words produced between Time 2 and Time 3. The reason for the increase is likely to be that immersion in an L2 environment provided greater opportunities for language learning on the one hand but also perhaps provided sufficient stimulus to reactivate vocabulary already known to learners on the other. This aligns with Meara's (2005) description of spontaneous vocabulary reactivation which takes place as a result of a short period of immersion in an L2 environment.

Thirdly, the gradual fall in words produced between Time 3 and 4 might be accounted for through the process of attrition. The delayed post-test (Time 4) seemed to reveal a decline or attrition in vocabulary knowledge after the SA experience had taken place. This experiment using Lex30, in common with previous studies which have looked at learners' vocabulary knowledge attrition (Weltens et al., 1989; Hansen et al., 2002; Schmitt, 2010; Ecke & Hall, 2012), showed clear signs of this. Although in this case it was not possible to assess if learners with larger vocabularies might retain their knowledge more effectively over time, future studies might well be able to achieve this. The evidence does seem to show that SA learners, on return to

their home L1 environment, experience some degree of vocabulary loss although it does not retreat to the level it was before SA.

4.5.2 Percentage v. raw scoring systems

In the experiment described in this chapter, scoring has been carried out by simply using the number of infrequent words that each participant produced. This method of scoring the Lex30 task has been mentioned in the previous chapter (see section 3.5.5). Raw scores represent the total number of infrequent (1K+) words that each SA participant produces at each time point. Some previous research (Fitzpatrick, 2003; Fitzpatrick & Meara, 2004; Clenton, 2010; Fitzpatrick & Clenton, 2017) has either used or considered the use of a percentage scoring system. Percentage scoring is where the Lex30 score represents the number of infrequent words produced as a percentage of all words produced. Fitzpatrick (2003) starts out by arguing that by using such a system it is possible to focus purely on word frequency as an indicator of the contents of a lexicon rather than the number of words. She notes the very large variation in the size of corpora that are produced by subjects, noting that the percentage scoring system is capable of negating any effect this might have on the Lex30 score.

However, at the same time Fitzpatrick (2003) also finds that the percentage system cannot always identify differences between individual test takers and changes in the performance of an individual test takers between two different time points. As previously mentioned in chapter 3 (see section 3.2.3), she carries out longitudinal studies on two groups (N =19 and N=16) of SA participants in an attempt to find if there is any change in Lex30 performance before and after an SA programme. She found that when the percentage scoring system is used there seems to be no significant difference between the mean scores of the subjects at test Times 1 and 2 ($p = .213$ and $p = .254$ respectively). However, when the raw score system was applied to the same data, there are significant differences between test Time 1 and 2 mean scores for study group one ($p = .029$) and also for study group two ($p = .018$).

In my experiment I found similar differences in the performance of both scoring systems. Between test Times 2 and 3 (where ANOVA repeated measures analysis of raw scores indicate that the largest changes take place) there was no significant difference when using a t-test on percentage scores ($p = .633$) while the

results expressed in terms of raw score show a significant change in SA participants' performance between test times ($p < .001$).

The percentage system takes no account of how many high-frequency words are produced. However, both Fitzpatrick (2003) and Clenton (2010) consider reasons why some participants in their studies also might produce more infrequent words and more words in general in response to the test task than others:

- (1) There is always some variation in the size of lexical pool or resource from which test takers select productive vocabulary items.
- (2) Subjects with a faster writing speed might be able to provide a greater number of words in response to Lex30 than those with a slower writing speed.
- (3) Some subjects are able to make a greater number and quicker connections between lexical items than others.
- (4) Motivational factors will mean that some subjects will perform differently.

The raw score is likely to include information about these four factors, but it is difficult to determine how they might interact when participants complete the Lex30 task.

In conclusion, it appears that both the percentage and raw score system have some merit. Which one is used will depend on the sort of information we are looking for. For a simple indication of the contents of the lexicon, the Lex30 percentage score provides information which might remain unaffected by one or more of the four features of lexical ability described above. On the other hand, precisely because it does include information about same factors, the Lex30 raw score might give a better indication of overall lexical proficiency. In this study I am interested in factors additional to vocabulary size so an increased capacity to produce words of whichever frequency - which might be linked to speed of access, confidence, fluency - is also of considerable interest.

4.5.3 Adopting a fine-grained approach (Kremmel 2016)

Unlike other frequency-based tests (the Vocabulary Levels Test; Nation, 1983; Schmitt et al., 2001), the Productive Vocabulary Levels Test (Laufer & Nation, 1999), Meara's yes/no vocabulary size tests (e.g., Meara & Buxton, 1987; Meara & Milton, 2002), Lex30 does not make any finer distinction between bands other than the 1000 marker. This is partly justified by the fact that with a maximum corpus size of 120, finer category distinctions would not be meaningful, and research such as that by Aizawa (2006) suggests that frequency band distinctions beyond about 4000 words are in any case not useful. Looking carefully at our results, adopting a fine-grained approach does not appear to reveal much further information. Higher frequency bands display a similar pattern where a rise between Times 1 and 2 is followed by a much steeper rise to Time 3 before experiencing a slight decline at Time 4. The fact that lower frequency word bands after the 2000 mark do not seem to follow a similar pattern of improvement seems to confirm Aizawa's findings that lower word frequency bands (particularly after 3-4000 words) lack usefulness. Aizawa (2006) tested 350 Japanese learners using a yes/no vocabulary test assessing vocabulary knowledge in terms of frequency bands. In a similar way to our own results he found a steady decline in knowledge within the first few higher frequency bands. As he progressed toward to lower frequency values this decline increases and making distinctions between words becomes less informative. A study by Milton and Alexiou (2009) found that a frequency model worked well on Greek EFL learners and could distinguish between the first four 1,000 frequency levels. Beyond this 4000-word threshold, however, it becomes more difficult to make any useful comparisons. With the data gathered in our experiment it seems that as words become less frequent, learners are less likely to acquire them in a strict predetermined order. A careful look at Figure 4.5 (p. 93) shows that SA participants become more unpredictable in their ability to produce words according to frequency, particularly beyond the 3000-word level. Before this threshold is reached there appears to be some degree of consistency within each successive frequency band – with the graph tracing a familiar pattern of a small followed by a greater increase in the number of words participants are able to produce up until Time 3 followed by a decline towards Time 4. SA participants are likely to be able acquire and produce words roughly according to their frequency during earlier stages in the learning process but this ability becomes less predictable the further we proceed along the

frequency continuum. It must be noted that this is the case at least for the lower proficiency participants in this study but for learners of higher proficiency levels, this finer grained approach to frequency analysis may yield different results.

4.5.4 Lex30: Possible future changes

Although Lex30 seems, on balance, a suitable measurement tool for this experiment there is case to be made for ways to further improve some of its features. The first point might include looking once again at the selection of cue words so that they might attract a better balance of responses. In our experiment cues like *cloth*, *fruit* and *window* encouraged test-takers to provide a large number of associate words while others like *dig* and *obey* attracted far fewer responses perhaps because they are more poorly understood. Other cues like *map*, *television* and *rice* seemed to attract a high number of words which then had to be discounted. These included, for example, words like *Google* (proper name) and *TV*, *CD* or *DVD* (acronyms). Perhaps refining the process by which such cues are selected might help provide more consistent results. A second point is improving the way in which Lex30 is scored. This includes deciding which response words to include and which to discount. Sometimes with past uses of Lex30 misspelled words, for example, have not been included or certain categories been given lower scores (e.g., Jiménez Catalán & Moreno Espinosa, 2005, p. 41). A third point might be to develop additional features for Lex30, for example, in effort to see how well the words that participants are able to produce in a Lex30 task are actually known (like Walters, 2012) or take Kremmel's (2016) methodology yet further and continue to tease out high frequency clusters at the higher frequency end of the frequency continuum into narrower bands. In chapter 8 I will consider such potential improvements in more detail.

4.6 Conclusion

This chapter has shown that Lex30 is capable of detecting significant changes in raw scores before and after an SA experience which agrees with a number of other studies (e.g., Fitzpatrick, 2003; Fitzpatrick & Clenton, 2010; Fitzpatrick, 2012). It has also revealed that interesting, although not statistically significant changes, can occur with participants' vocabulary knowledge in the preparation stages of an SA programme and in the period immediately following arrival back in their home country. Generally, findings show that the Lex30 test can be a useful and valid test of

productive vocabulary knowledge, allowing test takers to demonstrate the breadth of their vocabulary knowledge without constraint within a short time. However, some concerns about Lex30 administration, scoring and interpretation remain. The experiment has also shown that finer-grained analysis of words which occur towards the higher frequency end of the frequency continuum, using the narrower band widths suggested by Kremmel (2016) may help guide future decisions about pedagogy as well as SA programme design and preparation.

A number of questions still remain. These include identifying the different forms of vocabulary knowledge that are acquired or lost at each time point in our longitudinal study and trying to unearth evidence of the mechanisms at work which are responsible for their change. Some early signs of this, particularly with orthographic development, have already been indicated in this chapter. A study by Fitzpatrick (2012) examined the vocabulary development of a single subject over a period of nine months using Lex30. Using vocabulary samples that were gathered at regular time points during this time she found evidence, other than word frequency, which suggested that improvements, particularly in orthographical and collocational proficiency, may also occur. Based on this and several other studies, a more detailed investigation into these areas, as well as an examination into further changes in semantic knowledge and semantic diversity will be reported in the following chapters.

Chapter 5: Collocational changes during SA

This chapter looks at changes in collocational knowledge which may take place during an SA experience. It will start by underlining the importance of word combinations known as collocations as a measure of proficient language use. Attempts at defining and measuring collocations will be considered before reviewing relevant studies in past SA programmes. The second part of the chapter will present a new analysis of the data collected for the longitudinal study described in chapter 4. Questions about (1) whether SA might have any impact on collocational knowledge and (2) which proficiency level might be most affected will be addressed.

Data gathered using the Lex30 productive vocabulary task can lend itself to analyses beyond frequency-based scores. Investigating whether cue words used in a Word Association Task (WAT) can attract responses which also happen to be collocates may help us to understand if there are incremental changes in learners' knowledge of collocation during the course of an SA experience (Gobert, 2007; Fitzpatrick, 2012; Alqarni, 2017). Collocations (e.g., *black coffee*, *weak tea*, *terrorist attack*, *healthy lifestyle*) are one important type of formulaic sequence that has received increased attention in recent years as it has been argued that their study is a worthwhile activity with many potential benefits for learners across all proficiency levels (Bahns & Eldaw 1993; Nesselhauf, 2003; 2005). Recent research has looked at measuring collocation knowledge and exploring the way learning activities can help in its acquisition (Boers & Lindstromberg, 2009). However, despite this increased interest, it is clear that much more needs to be done to fully incorporate collocations in EFL teaching and testing (Higuera García, 2017). This chapter will attempt to firstly, investigate the importance of collocations and how they fit into overall linguistic knowledge. Secondly, it will attempt to provide a definition of what collocations are and discuss the various means with which they can be measured. Thirdly, it will look at previous research on collocations and SA and show how collocational knowledge can change over time. Finally, the chapter will report an experiment tracing changes in collocational knowledge with Japanese subjects participating in a short-term SA programme in the UK.

5.1 Importance of collocations

Collocation is often described as one form of formulaic language with Wray (2002) describing it as “hovering at the edge of definitions of formulaicity” (p. 47). As such collocations may be viewed as being less fixed and more fluid than other types of formulaic sequence, being more about tendencies and preferences. Bonk (2000) examines the terms *collocation* and *formulaic speech* and suggests that the former is about links between words in the mental lexicon based on lexical and semantic characteristics while the latter is about chunked storage and psycholinguistic reality. For the purposes of this chapter collocations are viewed as being an intrinsic part of, and not separate from, formulaic language.

The correct use of word combinations known as collocations is now regarded as an essential element of proficient language use (e.g., Cowie, 1998; Siyanova & Schmitt, 2008; Wray, 2002). Schmitt (2010) describes collocation as “one of the most important types of contextualized word knowledge” (p. 229). There is a general consensus that it is collocation that makes native speakers' speech idiomatic, fluent and smooth but it is also what makes L2 learners' speech sound unnatural. There is much evidence to suggest that L2 learners have problems with collocation particularly in their written and spoken language. Studies on collocation have shown that even high-level learners seem to experience problems in relation to using and developing L2 collocational knowledge (e.g., Arnaud & Savignon, 1997; Nesselhauf, 2005; Revier & Henriksen, 2006). Such evidence suggests that correct collocation use presents a major problem for L2 learners as it has been shown that collocational competence does not develop in parallel with general vocabulary knowledge (Bahn & Eldaw, 1993). As a result, learners often underuse or misuse native-like expressions and use unnatural word combinations instead.

Collocations are seen as highly significant not least because of their widespread occurrence throughout language. Conklin and Schmitt (2012) examined a range of different studies of texts which attempted to provide an estimate of the proportion of formulaic language within discourse. They found that these estimates ranged from one third to one half of the total content. Dealing more specifically with collocations, Howarth (1998a) looked at a corpus of L1 academic writing and found that 38% of word combinations, which used frequently occurring verbs together with nouns, were restricted collocations or idioms. Biber et al. (1999) reported that multi-word phrases constituted 28% of the spoken and 20% of the written discourse

analyzed. Following on from this, Siyanova and Schmitt (2008), using a similar kind of language corpus, found that more than 50% of adjective and noun combinations were collocations when compared to their usual frequency of occurrence in the British National Corpus (BNC). Although these figures may follow slightly different patterns it does appear that collocations account for a significant proportion of language discourse and have therefore attracted considerable interest from researchers.

5.1.1 The function of collocations

As well as their ubiquity in language, collocations serve a number of important functions. According to Henriksen (2012) they can help learners process language more fluently and having a knowledge of reliable collocational language ‘chunks’ can free cognitive capacity for other tasks. She also suggests that collocations can help language users differentiate between different polysemous words. Wray (2002; 2009) provides further evidence for the importance of collocations by saying that they can assist communication in a number of important ways. For example, collocations can help decrease the cognitive processing effort for speakers and also serve to minimize misunderstandings between them. She takes a similar view to Nesselau (2005) and Obukadeta (2019), in finding that collocations can be indicative of a shared community. This where they can sometimes come to reflect the sociolinguistic reality of language use in a particular group or context and which may be distinguished from other settings.

Collocations are also interesting to investigate as they can be useful for all language learning proficiency levels. Learners, even at the very earliest stages, seem to have at least some knowledge of collocations. For example, expressions like *make a mistake, strong coffee, or heavy rain*, are evident in elementary learners’ speech and writing. According to Peters (2016) the comprehension of some collocations may not be so challenging as producing them. Many meanings are transparent and can be understood with little difficulty when viewed in context. Peters goes on to provide empirical evidence of this where she compares both receptive and productive measures of collocation knowledge and concludes that “it is productive knowledge where learners’ main difficulties with collocations lie” (2016, p. 135). She notes that, although learners belonging to all proficiency levels seem to possess some degree of collocational knowledge it is also apparent that there is an important difference

between their simple comprehension and the ability to produce them in speech and writing in an appropriate manner. This raises some important questions about how knowledge of collocations can be measured and how this knowledge can develop with increases in proficiency level.

5.1.2 Defining collocations

Collocations have also attracted considerable interest from researchers as they are associated with two out of the nine aspects of Nation's (2001) "aspects of word knowledge" framework. With the receptive use of collocations the framework describes this as "words or types of words which occur together with the key word" and with productive use as "words which must be used together with the key word" (2001, p. 27). Based on this outline Brown (2014) suggests that a core definition of collocation can be described in terms of "words occurring or used together" and that this might well be accepted by many researchers as a starting point for any investigation (2014, p. 124). Questions remain, however, about clearer definitions of collocation use as a look into the background literature reveals the existence of two separate approaches (Barfield & Gyllstad, 2009, p. 2).

The first of these, the so-called phraseological approach, defines the multi-word units which make up collocations in linguistic terms and seeks to distinguish between phraseological units and free combinations of words. According to Brown (2018) collocations are seen as positioned at a certain point along a sliding scale, being less restrictive in use than idioms but a good deal more restrictive than free word combinations. Howarth (1998a) elaborates further by dividing this scale into four categories: free combinations, restricted collocations, figurative idioms and pure idioms and gives examples of word combinations in case. With the free combination category (the example he gives is: *blow a trumpet*) words are fully substitutable. The restricted collocation category contains one word which is specialized and used in a limited context (*blow a fuse*). An example of a figurative idiom is *blow your own trumpet* and a pure idiom is *blow the gaff*. He suggests that these four categories are on a continuum of collocability and restrictedness.

The second, frequency-based approach, has become more dominant and is linked with research on corpus linguistics. Originally introduced by Firth (1951) and further developed by Sinclair (1991) this can be described as examining the frequency of occurrence of word combinations within a text. For the purpose of

identifying collocations, corpora along with specialist software are used to look at recurrent word sequences or count the reoccurrence of words which appear within a certain distance span from each other. Following studies by Sinclair et al. (1970) most researchers presently use a 4-word span from the principal or node word in their search for collocations. Further analysis using statistical measures such as t-scores can also be used to assess whether two or more words are likely to be associated with each other and the strength of that association. There are a number of issues with adopting the frequency-based approach, an important one being the occurrence of word combinations which are both statistically significant and also appear to be collocations when they are not. Often free word combinations seem like collocations because of the subject matter in the text or where component words appear in such frequency with each other where in fact no discernable relationship exists. This issue is of such concern that some researchers (including Howarth, 1998a) have introduced additional statistical corrective measures when the frequency of component words exceed certain levels. Another issue is that application of the frequency-based approach to identify collocations is reliant on a number of decisions by the researcher including the degree of word span from the original node word used and the choice of statistical analysis.

Although corpus-based research has encouraged frequency-based approaches and has been more widely adopted, aspects of the phraseological approach are still worth considering. Recent research studies have tended to adopt both approaches using frequency-based measures to initially identify collocations and then phraseological methods used to further classify and characterize them. Combining both approaches recognizes the importance of the growing number of different corpora while also understanding that phraseology can reveal factors which are no less significant (Howarth, 1998b). Wray (2009) advises us that there does not need to be universal agreement on a one particular approach to collocation research or a single definition of what it means. She also warns that researchers should be aware of the implications of the definition which they eventually decide on and use in a particular case and to be careful to avoid using a definition which is only convenient because it fits the methodology that is being used.

5.1.3 Measuring collocations

Difficulties encountered with the definition of collocations are also reflected in the number of challenges with their measurement. A number of different tools, which vary in their complexity, can serve to measure collocation. Some researchers use strength of association measures (such as t-scores) to analyze L2 learner output while others have adopted a narrower approach by selecting targeted collocations. Bahn and Eldaw (1993) tested German speakers' translation skills using L1 German equivalents of target collocations with the requirement to translate each expression into English. Although this test seems like an effective way to measure learners' knowledge of collocations there are a number of problems associated with it including the fact that there are usually a number of ways to carry out a specific translation. The absence of the targeted collocation in the resulting translation might not be as the result of a total lack of knowledge and there is also the possibility that a sentence can be translated in a specific way that so that there is no requirement for a collocation in the first place.

Farghal and Obiedat (1995) used a cloze technique to study the collocation knowledge of Arabic speaking EFL learners. Using a number of carefully constructed cloze sentences they attempted to elicit targeted collocations such as "heavy drinker" in the following example:

George is a moderate drinker, but Peter is a _____ drinker

In a similar way to using translation to assess collocation knowledge there is a risk that the test taker will fill the blank with an alternative acceptable and not the targeted collocation such as the word "light" in the above example. One way around this might be to provide the initial letter 'h' to prompt the correct choice. Problems remain, however, as the cloze method only requires the participant to produce one element of a collocational pair and not the entire expression. Bonk (2000) takes a more balanced approach to collocation testing by producing three different formats including two kinds of cloze test and a multiple choice test. The first cloze test focuses on verb + object collocation, the second on verb + preposition and the multiple choice test requires the participant to identify incorrect example of collocation use out of four options. The results of Bonk's analysis show that the verb + object cloze test and the multiple choice test are valid methods of testing

collocation knowledge. Nesselhauf's (2005) extensive and detailed study uses a large learner corpus (154,191 words) consisting of 318 essays written by German L1 learners of English. Using the phraseological approach described earlier he identifies 2,082 instances where what are deemed "acceptable" collocations occur. The final decision whether collocation-like word combinations are actually collocations is made using four separate dictionaries based on the British National Corpus (BNC). The use of such dictionaries to clarify and simplify the identification of collocations is of particular importance to this thesis and are discussed in more detail in the next section.

5.1.4 The use of dictionaries to identify collocations

Brown (2014) gives an example of the previously described phraseological and frequency-based approaches with two specialist dictionaries used to identify collocations. The first is *The BBI Combinatory Dictionary of English* (Benson et al., 2009) which shows words along with their collocations which are separated into distinctive grammatical and lexical categories. The second: *A Frequency Dictionary of Contemporary American English* (Davies & Gardner, 2010) contains words based on their frequency of occurrence within a very large corpus. For each word listed there is a sublist of its collocates which are selected on the basis of how strongly they are associated. This strength of association is expressed as a statistical measure known as the Mutual Information (MI) value. Brown notes that it is, in fact, possible to combine the two approaches (2014, p. 125) as the second frequency-based dictionary mentioned (Davies & Gardner, 2010) uses frequency based measures to select collocations for inclusion but categorizes the information by part of speech.

A third dictionary must be mentioned here which displays much of the information above in a simpler and clearer way: *The Oxford Collocations Dictionary For Students of English* (Macintosh et al., 2009). This dictionary relies on the Oxford English Corpus as its primary source and is intended for productive use, most typically for help with writing. The dictionary includes the most frequent and useful British and American collocates for 9000 headwords. The selection criteria for item inclusion does not appear to be as rigorous as those used for other dictionaries and the collocations are not given in order of frequency of occurrence. But the great advantage is that, in most cases, a much greater choice of collocations for each headword is given which increases its value for learners. For some researchers and

particularly for the purposes of this thesis this dictionary can help with analysis. Barfield (2009) used it to draw up list of collocates to measure Japanese L1 EFL learners' collocation knowledge in a study replicated by Brown (2018). With the use of the dictionaries described above researchers are not required to make any decisions with criteria for the make-up of the lists having already been selected by the dictionary editors. The *Oxford Collocations Dictionary* shows that it was compiled in a similar way, but the exact process of compilation is not completely explained. It mentions that "the main source" (p.viii) of data was a corpus (Oxford English Corpus) and that compilers were able to "check how frequently any given combination occurred, in how many (and what kind of) sources, and in what particular contexts" (p. viii). There might be a concern, therefore, about how the same compilers can make principled decisions in this area. Davies and Gardner (2010) in their dictionary admit that using some forms of measurement to identify collocations "is sometimes more an art than a science" (p. 6). It seems that some concern must remain in using specialist collocational dictionaries since some of their content may still depend on the occasional arbitrary choice made by their creators.

5.2 Collocation and study abroad

Few studies have looked the SA experience and how an L1 environment may affect the acquisition of collocation knowledge. This section will look at research by Fitzpatrick (2012), Gobert (2007), Alsakran (2011), Alqarni (2017) and others which has explored the impact that SA may have in terms of the proficiency level of the participants, the length of study overseas and the influence that an L1 may have on the production of collocations. It will go on to assess the importance of some of these factors in designing an experiment to measure changes in collocation knowledge of Japanese subjects participating in a short-term SA programme.

5.2.1 Acquisition of collocation knowledge

Fitzpatrick's (2012) study looks at the impact of SA and finds that it can increase growth in overall vocabulary knowledge in the case of a single subject . She suggests, however, that details of the nature of this growth are unclear and it is unknown whether SA favours certain aspects of vocabulary acquisition over others. Fitzpatrick's case study investigates a single Chinese L1 speaker studying on a university course in the UK. The subject is tested at 6-week intervals over the course

of an academic year. Using data collected by the Lex30 productive vocabulary task administered at each time point, she examines changes in a number of features including orthographic and morphological knowledge and form-meaning links. Importantly, for the purposes of this chapter, collocation is also considered and both Fitzpatrick's experiment and the present study look at the relationship between Lex30's cue words and the responses given by the subject. Fitzpatrick found that the vast majority of responses (91%) indicated a preference for meaning-based (paradigmatic) associations as opposed to form-based (clang) associations (0%) or position-based (syntagmatic) associations (9%). In Word Association research, collocations are typically classed under the "syntagmatic" or "position-based" category which, in this case, showed a small but detectable rise in patterns of collocation use.

Other research which has relevance to SA and collocational knowledge has also been conducted. Gobert's (2007) qualitative study reported on the low L2 English collocational knowledge levels of 29 advanced Saudi students studying At Home (AH) in an L1 environment despite explicit instructions on collocation being taught in the classroom. Gobert concluded that this showed that collocation knowledge was perhaps acquired rather than learned and that one crucial factor could be exposure to target language. She concluded that this would explain why even low proficiency learners studying abroad may in fact have a better collocational knowledge than advanced learners remaining at home. Alsakran (2011) also provided evidence of the beneficial effects of an L2 environment. Her study looked at the productive and receptive knowledge of lexical collocation knowledge of Arabic L1 speakers. What made this study interesting was the comparison between ESL learner (SA) and EFL learner (AH) groups. The results of the study revealed that the L2 learning environment seemed to have a strong influence on the acquisition of L2 collocations and that the ESL SA learner group scored significantly better than the AH one. Amed (2011) in her qualitative study of 60 Saudi Arabian students in Australia, noted the enhanced process of language acquisition of 60 Saudi Arabian students in an English-speaking environment. Her study found greater improvements with students who had been encouraged to study abroad on previous occasions and that much language improvement was due to home stays and direct native speaker contact. The study helps identify a number of factors which can

enhance language acquisition in a L2 environment. In particular, experience of homestays seemed to be an important criterion.

Alqarni's (2017) study investigated the impact that SA has on the lexical knowledge of Saudi Arabian students studying in Australia. He focused on two questions: firstly, whether the length of SA had a significant effect on the acquisition of lexical collocations and secondly, whether there is a significant gender difference in the acquisition of this knowledge. He conducted a cross-sectional study using a multi-choice Collocation Test (CT) of 12 questions with one point awarded for each correct answer. The test he used was adapted from an earlier test used in Gobert's (2007) study. He also used a demographic questionnaire to collect data on students' length of stay and experiences in Australia. He conducted the test on 124 (male: 92 and female: 32) Saudi Arabian students studying at a number of academic institutions across Australia from a period of between one and four years. Results showed that there was a positive correlation between the length of SA and knowledge of collocation. The mean scores of the CT increased the longer the SA participants spent in Australia. It was also found that there was no significant difference between male and female scores. The most marked increase was during the first year (mean scores were 62% on the CT). Alqarni compared these scores to those gained by advanced EFL students studying in an AH environment as reported by other researchers (Gobert, 2007; Alaskran, 2011; Shehata, 2008). Participants who had stayed two or three years in Australia, showed minimal improvement (63.6% and 64% respectively) but those in the four-year group showed a score of 73%. A t-test revealed that the difference between the one-year group and the four-year group was the only significant one (11%). Results confirm earlier studies by Alaskran (2011), Amed (2011), Milton and Meara (1995) and Storch and Hill (2008) and support the positive effects of the length of stay in a L2 learning environment on lexical collocation knowledge. Based on this research, it is likely that SA can improve such knowledge in ways which cannot be achieved in the home country.

5.2.2 Influence of culture and first language

A related issue is the influence of culture and first language on collocation. Ooi et al. (2007) provides an example of the unique nature of Singaporean English referring to the collocation *weekend car* (referring to a car that can only be legally driven outside of peak commuter times). With Japanese L1 speakers the production

of English collocates may be influenced by both culture and language. A considerable number of Japanese words are based on foreign language terms including *wasi-eigo* (Japanese pseudo-anglicisms) due to a large number of western concepts imported during Japan's modernization in the Meiji period (post-1868). Some of the words have evolved into frequently occurring combinations or collocations the meaning of which may differ to those imagined by native speakers. Examples would include the term *baby car* which does not refer to a small sized-car but actually a baby's pram or pushchair. *Seat knock* refers to a term used in baseball fielding practice while *silver seat* does not refer to a seat which is painted silver but rather a priority seat for elderly people on a bus or train. This might be of some importance when Japanese L1 EFL learners' production of English collocations is evaluated, since there appears to be a cultural element in some learners' collocational production and acceptability. Any mismatch between the cultural basis of the production and whether they are acceptable in English might become problematic and perhaps distort the full picture of learners' knowledge of collocations.

5.2.3 Towards a new experiment

The main purpose of the experiment described in this chapter is to examine if SA has any impact on the acquisition of lexical collocations. Past evidence from studies with Saudi Arabian students (Gobert, 2007; Shehata, 2008; Alsakran, 2011; Amed, 2011 and Alqarni, 2017) and a single case study (Fitzpatrick, 2012) reveal that significant changes in collocational knowledge can occur over time perhaps as a result of exposure to an L2 environment. There are a number of differences in the direction that a new study might take:

- (1) Most of the studies described above were cross-sectional in their methodology. In other words, information and test scores were gathered under two different conditions (i.e. study abroad and study at home). A new study could be longitudinal in nature (similar with Fitzpatrick 2012) with experimental data being collected at a number of time points and then collocations produced by the same group participants compared.
- (2) The profile of the experimental subjects is likely to be different. Alqarni (2017) described his subjects as 124 Saudi Arabian students (with 92 males and 32

females), most likely of young age (although he includes no specific ages in his study) following different SA programmes across Australia. Alqarni also makes no mention of L2 proficiency levels or any differences in individuals' particular L2 environments which could have affected language acquisition. The profile of the subjects in our new study has been described earlier in this thesis (see chapter 4). To restate briefly the study group comprises 38 female Japanese L1 participants aged between 18 and 19 studying at one institution in the UK and all staying with host families. In addition they are regularly engaged in similar out-of-class activities and to this extent there is some degree of confidence in the degree of uniformity of the participants' L2 learning environment.

(3) The length of study has some importance. Alqarni (2017) looked at students undergoing SA for periods ranging from one to four years while Fitzpatrick (2012) looked at changes occurring within a shorter time scale (eight months). The new study will seek to identify if changes in collocation knowledge can be detected within a period of less than one month.

(4) To test subjects' collocational knowledge, Alqarni (2017) used a receptive Collocation Test (CT) consisting of multiple choice questions. In the new experiment productive vocabulary knowledge, including collocations, will be tested using data collected by a word association task (Lex30). This is an identical task to the one used by Fitzpatrick (2012) although the way in which collocations are identified will differ. Use of an established reference source such as a dictionary is one way of checking whether word combinations are collocations. If collocation identification can be carried out in a reliable and consistent manner this might alleviate some concerns about the arbitrary nature about the process of selection.

One final point must be raised. Alqarni questioned why minimal growth occurred after an initial increase in the first year. He quotes mean CT scores rising from a mean score of 62% after one year of study to 63.6 % and 64% respectively in the second and third years. He suggested that this demonstrated that collocation knowledge needs some years to achieve significant growth rates after experiencing more promising results at the start. The evidence suggests that lower proficiency

learners may derive the greatest benefit from SA at least at the beginning stages of their SA programme and therefore most likely to make the most improvement. The scores of subjects in their third and fourth years of SA perhaps show that more proficient learners experience greater difficulties in demonstrating progress. What might be interesting to discover with a new study is if there is any difference between groups of relatively higher and lower proficiency learners in the numbers of collocations acquired during an SA experience.

5.3 Research questions

Taking into account the results and reviews of past research studies on a variety of learner groups with different proficiency levels, L1s and cultural backgrounds, two research questions can be proposed:

- (1) Does a learner's productive knowledge of L2 collocations change over a 15-day SA period?
- (2) Are certain learner groups, based on vocabulary proficiency level, more likely to increase production of collocations?

5.4 Experiment: Collocation and study abroad

The main aim of the experiment is to investigate if there are any changes in productive collocational knowledge of 38 Japanese L1 participants before, during and after a short-term 15-day SA experience in the UK. As described earlier in this thesis (Chapter 4) a word association task (Lex30) was used to gather a sample of productive vocabulary on four separate occasions 23 days before departure, just before departure to the UK, two days day after return and then 21 days after arrival back in the students' home country.

The 38 subjects are divided into two separate groups on the basis of approximate language proficiency. It was especially difficult to make an accurate assessment of SA participants' proficiency level. Nakamura University has no formal language testing policy whereby students are assessed at the beginning and end of their course. A possible reason for this is the fact that they are not studying English language as a core part of their university studies.

English proficiency tests are widely available in Japan. The most common of these is the Nihon Eigo Kentei Kyokai (commonly referred to as Eiken) proficiency test (see section 2.4.2). It consists of a total of seven levels, from Grade 5 to Grade 1, including Grade Pre-2 and Grade Pre-1. There is some evidence that Eiken compares well with other, more internationally known exams (TOEIC, TOEFL, IELTS) (Hill, 2010, In'nami & Koizumi, 2017).). It has also been demonstrated that intermediate levels of the Eiken are comparable with CEFR measures (Fujita, Yokouchi, Matsuoka, Nakamura & Hirai, 2016). Unfortunately, partly due to their non-language areas of study, there is evidence that only a limited number of our SA participants (less than half) have taken Eiken at any time. Even fewer students have taken TOEIC, TOEFL and IELTS especially as they tend to be more expensive. In recent years there has been some concern that students' financial conditions or residential location is having some impact on their ability to take such tests. (Kuwabara, 2021).

Due to the fact that few SA participants had taken any form of established proficiency test (such as TOEIC) it was decided to use a more approximate means to measure language ability. The total number of infrequent words (1K+) that the subjects had produced during their four test attempts was taken as a single raw score and would serve as a rough indicator of subjects' proficiency level for the purposes of the experiment. Fitzpatrick and Clenton (2010) found that scores, obtained by using Lex30, could serve as an approximate guide to vocabulary proficiency level as they correlated significantly with other vocabulary measures including the Productive Levels Test (PLT) (Laufer & Nation, 1999) and a translation test. Fitzpatrick (2007) clearly argues the need to be cautious when using such a method of test comparison citing that a major difference between the tests is that Lex30 uses no predetermined test items and correlates more weakly with the other two than they do with each other (2007, pp. 127–128). It is with some degree of caution then that such a method of distinguishing between different levels of ability is used in this experiment.

5.4.1 Method of analysis

The Lex30 uses a word association task to extract a sample of productive vocabulary using 30 specially selected cue words which are used to elicit up to four responses per words. The response words produced were carefully examined using a list of collocates derived from *The Oxford Collocations Dictionary For Students of*

English (Macintosh et al., 2009). For every word which matched a collocate listed in the dictionary a point was awarded. In this way a score was calculated for each subject which was the total number of collocates that they were able to produce at each of the four time points. Some of the issues with the use of dictionaries in the identification of collocations have been previously discussed (see section 5.1.4) but the main advantage offered is that this dictionary provides a greater choice of collocations for each headword given. It is important to underline the fact that at no time were subjects asked to produce collocates in response to cue words but it is probable that in many cases they would be likely to do so.

5.4.2 Results: Overall Performance (RQ1)

Table 5.1 shows the same information for the group as a whole as on the graph. The mean number of collocations produced by the 38 participants increased from test Time 1 to test Time 3 before decreasing slightly at test Time 4. The greatest increase (1.81) occurs between test Times 2 and 3.

Table 5.1

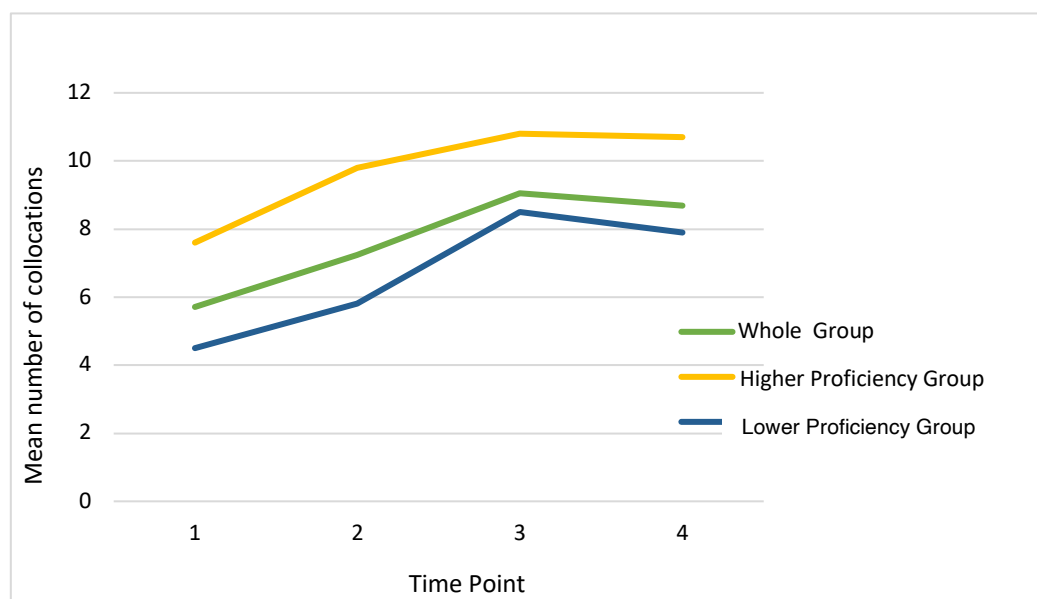
Descriptive Statistics: overall performance

	N	Mean	SD
Test Time 1	38	5.71	2.894
Test Time 2	38	7.24	3.044
Test Time 3	38	9.05	3.654
Test Time 4	38	8.68	3.129

Figure 5.1 shows a colour graph depicting the mean number of collocations produced per participant across a range of proficiency levels. Looking at the green “whole group” line it appears that there has been a general increase in the number of collocations produced by participants during the course of the SA programme. Between Times 1 to 3 the number increased steadily before experiencing a slight decline in the period after the participants’ return to Japan.

Figure 5.1

Mean number of collocations produced per participant



To find out whether these differences are significant a repeated measures ANOVA analysis was conducted (see Appendix 5). A test of within subject effects showed that mean collocation scores differed significantly across four time points: $[F(3, 111) = 16.782, p < .001]$. The effect size, calculated as eta squared (η^2), was 0.312, indicating a medium effect. A post hoc pairwise comparison using the Bonferroni correction gives more detailed results. It shows that a greater number of collocations are produced at time one than at time two (7.24 vs 5.71) and that this was statistically significant (-1.53 (95% CI - 2.83 to - 0.22) $p = .014$). From Time 2 to Time 3 there was an increase in the mean rate of collocation acquisition (7.24 vs 9.05) which was also statistically significant (- 1.82 (95% CI - 3.35 to - 0.29) $p = .013$). There was only a small difference in the number of collocations produced between Time 3 and Time 4. The number slightly fell (9.05 vs 8.68) but this decrease was not significant (.37 (95% CI - 1.02 to 1.75) $p = 1.00$). Looking more widely at the results - between Time 1 and Time 3 and between Time 1 and 4 the increase in collocations was significant (both comparisons indicating $p < .001$). Therefore, we can conclude that the results for the ANOVA indicate a significant increase in the number of collocations produced in the periods before and during the SA experience and that the slight decrease seen on the participants' return to their home country was

not significant. Overall between Time 1 and Times 3 and 4 there was a significant rise in the collocation use.

5.4.3 Variation by proficiency group (RQ2)

Table 5.2

Descriptive Statistics: Higher proficiency group

	N	Mean	SD
Test Time 1	19	7.05	3.064
Test Time 2	19	8.68	3.215
Test Time 3	19	10.21	3.809
Test Time 4	19	9.47	3.425

The entire group of 38 participants was then divided into half on the basis of proficiency level. This level was decided on the basis of the highest Lex30 raw score in terms of low frequency words produced during all four attempts during the longitudinal study. In the absence of any alternative means of language proficiency assessment this ‘rough and ready’ method might be enough to differentiate between the two groups in a meaningful way. Table 5.2 shows the results of the higher proficiency group now comprising of 19 subjects. These results were analyzed using repeated measures ANOVA (see Appendix 6) to find out whether the differences are significant. A test of within subject effects showed that mean collocation scores differed significantly across four time points: $[F(3, 54) = 6.305, p < .001]$. The effect size, calculated as eta squared (η^2), was 0.259, indicating a small effect. A post hoc pairwise comparison using the Bonferroni correction gives more detailed results. It shows that a greater number of collocations are produced between the Time 1 and the Time 2 (7.05 vs 8.68) and that this was not statistically significant (- 1.63 (95% CI - 3.87 to 0.60) $p = .265$). From Time 2 to Time 3 there was an increase in the mean rate of collocation acquisition (8.68 vs 10.21) which was also not statistically significant (- 1.53 (95% CI - 3.73 to 0.6759) $p = .328$). A small difference in the number of collocations is apparent between Time 3 and Time 4 with the number slightly decreasing (10.21 vs 9.47) but this decrease was not significant (0.421 (95% CI - 1.80 to 2.65) $p = 1.000$). Again looking more widely at the results - between

time 1 and time 3 the increase in collocation use was significant (- 3.16 (95% CI - 5.74 to - 0.572) $p = .012$) but between Times 1 and 4 the increase in collocation use was not significant (- 2.74 (95% CI - 5.15 to - 0.32) $p = .021$). Therefore, we can conclude that the results for the ANOVA indicate that the only significant increase in the number of collocations produced by more proficient students was between Time 1 and Time 3, a period starting 23 days before and ending just after completion of an SA experience.

Table 5.3

Descriptive Statistics: Lower proficiency group

	N	Mean	SD
Test Time 1	19	4.37	2.006
Test Time 2	19	5.79	2.070
Test Time 3	19	7.87	3.178
Test Time 4	19	7.59	2.411

Table 5.3 shows the descriptive statistics for the lower proficiency group. To find out whether these differences are significant a repeated measures ANOVA analysis was again conducted (see Appendix 7). A test of within subject effects showed that mean collocation scores differed significantly across four time points: $[F(3, 54) = 10.794, p < .001]$. The effect size, calculated as eta squared (η^2), was 0.375, indicating a medium effect. A post hoc pairwise comparison using the Bonferroni correction gives more detailed results. It shows that the mean number of collocations produced between the Time 1 and Time 2 (4.37 vs 5.79) increased and that this was not statistically significant (- 1.42 (95% CI - 3.12 to 0.279) $p = .141$). From Time 2 to Time 3 there was another increase in the mean rate of collocation acquisition (5.79 vs 7.89) which was also not statistically significant (- 2.11 (95% CI - 4.54 to 0.33) $p = .119$). There was a small decrease in the number of collocations produced between Time 3 and Time 4 (7.89 vs 7.58). This decrease was not significant (0.316 (95% CI - 1.67 to 2.31) $p = 1.00$). Although the results appear to show that there was no significant change between subsequent time points this is not the case over the longer term. Between Time 1 and Time 3 and Time 1 and Time 4 there were significant

changes in collocation use. These were (-3.53 (95% CI - 6.06 and - 0.99) $p = .004$) and (95% CI - 4.89 to - 1.55) $p < .0019$) respectively.

5.5 Discussion

Firstly, in answer to the two research questions there does appear to be an overall significant rise in collocational production over the period of SA which supports Fitzpatrick's (2012) findings. Figure 5.1 and repeated measures ANOVA analysis shows that the greatest increase takes place in the periods immediately before departure and during the SA experience itself. Individual participants seemed to produce a sufficient number of collocational responses to the cue words to demonstrate that some sort of change is taking place. In Chapter 8 a number of ways in which the number of collocational responses can be further encouraged will be discussed.

Secondly, lower proficiency learners, as defined by a "rough and ready" system of Lex30 scores, seem to acquire collocational knowledge more easily compared with higher proficiency learners. Previous research by Alqarni (2017) suggests that lower proficiency learners may derive the greatest benefit from SA at least at the beginning stages of their SA programme and are therefore most likely to make the most improvement. The scores of subjects committed to long-term SA (in their third and fourth years of an SA experience) perhaps show that more proficient learners have greater difficulties in demonstrating progress. Lower proficiency students make more progress (or at least students in their initial year of study). This, again, supports Alqarni's 2017 findings which demonstrate that collocation knowledge reaches a kind of plateau beyond which change becomes more difficult. In other words after experiencing promising results at the start of SA, a considerable time needs to pass before significant growth rates can once again be experienced. The evidence suggests that lower proficiency learners may derive the greatest benefit from SA at least at the beginning stages of their SA programme and therefore most likely to make the most improvement. The method whereby an experimental group is divided into two categories depending on proficiency has its own limitations and issues. Ideally, each SA participant would undergo a form of proficiency assessment separate from Lex30 task. However, given the fact that no independent assessment is made before the SA programme, making a judgment of an individual's proficiency based on Lex30 score was thought to be a viable but not completely reliable option.

In scrutinising the responses to identify collocational behaviour, a distinct cultural influence on response words was noted. With the cue words *attack* and *rice*, for instance, many of the responses that the test-takers provided were very different from the collocates listed in the dictionary. Instead of expected collocational responses like *surprise* and *fried* words like *volleyball* and *curry* indicating that L1 cultural influences can also be important in deciding which response may be most appropriate. With the cue word *attack*, many collocates listed in the dictionary are adjectives concerned with violence or war, or other examples like *strong* or *surprise*. In the experiment the most commonly produced response word was *volleyball*. A possible reason for this that ‘*Attack*’ is the name of a popular comic about volleyball in Japan. It is a popular sport in the country particularly with young women which is perhaps why the word appeared so often in response to the cue. With the cue word *rice*, culturally influenced responses such as *miso* or *curry* were produced instead of dictionary-listed collocates like *white* or *fried*. Clearly many of the responses given in the experiment were not L2 collocates (see *Defining collocations* section 5.1.2) but rather examples of the influence that L1 cultural or linguistic factors can have.

It is worth noting that many in the English language teaching world still treat English as though it were represented only by native-speaker models. Kirkpatrick (2007) argues against the adoption of a such models as a target for Chinese L1 learners of English saying that it can be demotivating for many students. Japanese L1 learners might also realise that such a model is likely to be unattainable by most and might become frustrated by setting themselves what is, in effect, an impossible target (Cook, 2001). In this experiment many responses supplied by SA participants were not collocations in the native-speaker sense but particular examples occur often enough over time to seem so. It seems as though further discussion is warranted concerning non-native speaker varieties of collocates which might lead toward some acceptance of their validity.

5.6 Conclusion

The results from the experiment support research by Fitzpatrick (2012), Alqarni (2017) and Gobert (2007) which suggest that SA can have a positive impact on collocational knowledge. The results also show that any improvement is more likely to take place with lower proficiency SA participants.

There are two possible improvements that might be made in any future experiment. The first is to find a more accurate method for assessing proficiency levels. This would allow a better understanding of how different proficiency groups can differ in their rate of collocational knowledge improvement. The second would be to modify the test design so as to encourage test-takers to provide a higher number of collocates. The present version of Lex30 uses only a limited range of cue words. Increasing both this number and the responses given to each cue would result in greater accuracy in any future assessment of collocational knowledge.

Chapter 6: Orthographic changes during SA

There is an belief associated with vocabulary knowledge, that words are either ‘known’ or ‘not known’, and that the former will gradually increase in number as learning progresses. However, in the interlanguage of the non-native speaker, the acquisition of word knowledge is more complicated and dynamic. Learners will have a good knowledge of some words but sometimes the knowledge of others may be unstable or incomplete. Such imperfectly known words are still of considerable interest to the researcher as they may help us to better understand the process of how new language is acquired. Tracking orthographic (spelling) changes during a participant’s Study Abroad (SA) experience may be one way of accomplishing this. Using “written form”, one of the nine aspects listed in Nation’s word knowledge framework (2001, p. 27)), this chapter presents an analysis of Japanese L1 English language learners’ word form knowledge. By examining the written responses to the cues in the Lex30 task at four test times, it may be possible to track incremental changes in SA participants’ orthographic accuracy and by doing so, make some assessment of the partial knowledge they may have of certain word forms.

As well as reviewing past research on orthographic development this chapter also looks at the influence that L1 language and culture can have on English spelling. The study reported in this chapter uses the same data sets from Japanese SA participants as the studies in chapters 4 and 5. Here, the data is scrutinised to determine the effect of the SA experience on spelling, whether certain proficiency levels are favoured, and whether Japanese language and culture might affect the way English words are spelled.

6.1 Orthography and depth of knowledge

As we have already established in previous chapters, many of the processes which occur in language acquisition during SA remain little known. This chapter steps away from the word frequency-based analyses addressed in previous chapters to make some observations about changes in patterns of orthography which may occur during a short-term SA programme. The focus on orthography is arrived at by considering it as a means with which to assess the degree of familiarity or depth of knowledge with which a particular lexical item may be known. It also draws on previous research (e.g., Schmitt, 1998a; Bell, 2009; Churchill, 2008; Zareva, 2012)

which shows that it is possible to measure partial vocabulary knowledge to reveal gradual changes which can be traced over time.

The chapter will go on to summarize a case study by Fitzpatrick (2012), previously mentioned in connection with collocation use, which also looks at orthographic differences that may indicate changes in the degree in which an individual learner develops mastery of certain lexical items. It will also investigate the question of whether certain patterns of orthographic change can give an insight into writing challenges connected with particular L1 groups, in this case Japanese EFL writers. After reviewing previous research, the chapter reports an analysis, partly inspired by Fitzpatrick's work, of data from 38 Japanese L1 EFL learners (the same dataset as was used in chapters 4 and 5) and examines the orthographic changes which occur during the course of their SA experience.

Research using a number of different measuring instruments has indicated that quantitative improvement in participants' language ability takes place during the SA experience (e.g., Milton & Meara, 1995; Parker, 1999; McManus et al., 2020). Many studies have focused on depth rather than breadth of vocabulary knowledge, often conceptualising depth as a number of different aspects of knowledge. This section will examine attempts to measure depth of knowledge and partial word knowledge, to give context for the focus on orthographic knowledge in the rest of this chapter.

Trembly (1966) made the observation that it is not only how many words are known by learners that is of importance but also how well those words are actually known. He proposed that certain words be known as "frontier words", or partially familiar words, and suggested that these "exist in the frontier region between the point where every word is known and the point where no words are known" (1966, p. 229). More recently other researchers including Wesche and Paribakht (1996), Read (1993), Schmitt (1998b), Bell (2009) and Churchill (2008) have also considered this "frontier region" and attempted to measure partial knowledge within it and how this might change over time. Paribakht and Wesche (1993) created a Vocabulary Knowledge Scale (VKS) to measure "certain states in the initial development of core knowledge of given words" (1993, p. 29). Using a word familiarity scale it examined L2 users' self-awareness in distinguishing between unknown, partially known, and familiar vocabulary, as well as into their ability to use the tested words in sentences of their own. Schmitt (1998b) used a wide range of

methods to observe the acquisition of a small number of pre-selected lexical items. Over the course of a year's study he examined the spelling, meaning, associations and grammatical characteristics of 11 target items, tracing gradually deeper levels of understanding, from purely receptive knowledge towards the point where the learner was capable of using items productively in a variety of contexts.

Bell (2009) and Churchill (2008) both adopted a similar approach to Schmitt using case studies to investigate stages in the language acquisition but with one important difference: the vocabulary items they examined were undecided at the start of each of their experiments. Churchill tracked his own three-month journey towards a gaining greater knowledge of a single item of Japanese vocabulary: *saiketsu* (a word meaning decision, ruling or judgment), recording in detail the nature of each encounter. The single target word presented itself for selection as it was repeatedly heard in a variety of linguistic contexts and the author became gradually more interested in its use. Similarly, Bell used 28 essays written by single Korean L1 ESL learner gradually narrowing down his search to concentrate on the developing use of 17 newly discovered lexical items over a 16-month period. The 17 items were selected because they tended to frequently reoccur throughout the data that the subject had produced. Both studies were small scale single case studies which successfully demonstrated that the process of vocabulary acquisition is multi-dimensional in nature. They also showed that a new focus on the "micro-development" (Churchill, 2008, p. 339), of individual lexical items can help bring about a better understanding about the nature of the processes at work.

6.2 Partial word knowledge and Zareva (2012)

In a discussion on depth of vocabulary knowledge measurement, mention should be made of work by Zareva (2012) who examined the degree to which partial word knowledge can occur within the lexicon. She compared three groups: one consisting of native speakers and two others with advanced and intermediate learners of English. She carried out her research within the framework of Richards' (1976) taxonomy of aspects of word knowledge which describes in some detail what actually knowing a word entails and has been recounted in various forms by a number of other researchers (e.g., Dale, 1965; Wesche & Paribakht, 1993; Nation, 1984; 2001). The seven aspects include:

1. knowing the probability of encountering a word in speech or writing;
2. knowing the limitations of word use according to function and situation;
3. knowing its syntactic properties;
4. knowing the word's underlying form as well as its derivations;
5. knowing the associations between the word and other words in the language;
6. knowing the semantic value of the word;
7. knowing many of the meanings associated with the word.

Zareva set out to examine the partial word knowledge within her participant groups with regard to several features from Richards' (1976) taxonomy including understanding a word's semantic value, having a knowledge of its syntactic properties, knowing which other words are likely to be associated with it and underlying word forms. The participants self-reported the degree of partial familiarity they had with a number of low and mid frequency content words using a word knowledge scale. Their self-assessment was later checked to demonstrate actual as opposed to perceived knowledge of the given target items. The results showed that there were three distinctive patterns of partially familiar vocabulary including knowledge of lexical class and word meaning. However, their distribution across the three proficiency groups was quite different, which indicated that partial knowledge was linked to different word features. Although the study investigated only a limited number of aspects of Richards' word knowledge framework it was, nonetheless, an important attempt to investigate the phenomenon of partial word knowledge which all language users experience, be it in their LI or L2, in the process of developing their lexicons.

To summarize, discussions about depth of vocabulary knowledge seem to indicate that for L2 learners knowing a word is not an either/or state of affairs. At one extreme knowing all the properties of a word implies that all 7 of Richards' criteria are satisfied and on the other end of the scale there are words which the learner will be completely unfamiliar with where none of the criteria are met. Between these two extremes there are a number of possible stages of partial knowledge. What does seem clear, however, is that there has been a decided shift away from the view that learners have only a binary knowledge of words (known/unknown) towards an appreciation of the fact that there are degrees of familiarity of that knowledge.

6.2.1 Measuring orthographic changes

Tracing changes in patterns of learner spelling may offer a fresh way to examine the development of word knowledge. Aspects of word knowledge, according to Richards (1976) include the ability to understand words and reproduce them accurately with correct spelling. Tracing the gradual development of individual word spelling from the earliest stages of initial comprehension to higher levels of learner competence where words can be reproduced accurately and confidently may offer new insights into the complex processes of vocabulary acquisition. There is another view, however, that this gradual process towards complete comprehension may not be so straightforward. Partial knowledge may not necessarily remain as a stage in the progression towards a fuller knowledge but rather become a legitimate 'state' in its own right because it is adequate for a particular learner's needs.

There may be some concerns with measuring changes in this way. The English orthographic system contains a number of inconsistencies and often confounds L2 and L1 speakers alike. The spelling of words has been described as "a rag-bag of lawlessness" due to the complicated relationship between the spoken and written word (Wistor, 1907, p. 35). Crystal (1992) suggests that this perception is partly due to the "400 or so irregular spellings which are largely among the most frequently occurring words in the language" (p. 214). Van Berkel (2005) goes on to explain that English, with its 44 phonemes representing the 26 letters of the alphabet, must be considered orthographically deep as the connection between sound and symbols may be interpreted in a number of different ways. The consensus is that written English is indeed difficult but there is the view that good spelling is nonetheless a sign of a good education and for L1 speakers at least, should be seen as a crucial factor for presenting themselves educationally and professionally (Horn & Ashbaugh, 1920). With L2 users of English effective spelling may be just as important not least because it helps with better communication. However, L2 users and their performance with English spelling has not been covered so extensively by the established research although specific language groups, and their associated problems, have been looked at. Examples of research which explore the English spelling of L2 users include Martínez Adrián and Gallardo del Puerto (2017) with Spanish, Ibrahim (1978) and Alenazi (2018) with Arabic, Wang and Geva (2003) with Chinese and Mark (1998) and Okada (2002; 2004) with Japanese L1 learners and speakers.

The challenge of mastering English orthography is a difficult one, particularly in the case of L2 learners, where the journey from first hearing a new word, to reading it in a text, to a point where the same word can be reproduced accurately in context, might be long. However, it could also be an effective way of gaining a new insight into the stages of vocabulary development. I will start looking at research by Fitzpatrick (2012) which shows that orthographic changes can occur over time but that this may not always necessarily indicate improvements in spelling accuracy. Then a study by Cook (1997) will be considered which describes how the spelling of English words can be affected by the relationship between the sounds and spelling of L2 learners' original native language. To this end, Okada (2002; 2004) and Gunion (2012) look specifically at the relationship between the Japanese writing system and whether it may have any influence on Japanese EFL writers particularly during an SA experience. Finally, research by Tuladhar and Akatsuka (2017) will be considered to see if it may be worthwhile examining whether learning accurate pronunciation through speaking activities, a focus of many SA programmes, can actually help improve spelling accuracy.

6.2.2 Fitzpatrick (2012) and tracking orthographic changes

Fitzpatrick (2012) used a Word Association Task (WAT) to elicit samples from a single subject during a period of SA. The WAT used the cue words from Lex30 (Meara & Fitzpatrick, 2000), which has been described and discussed earlier in this thesis. Fitzpatrick argued that an important advantage of a WAT, in a similar way as the experiments carried by Bell (2009) and Churchill (2008), is that it focuses on lexical items produced by learners in free and spontaneous matter. By avoiding the use of predetermined targets or answers which are used in some earlier studies the "elicitation process should result in minimal contextual interference or scaffolding" (Fitzpatrick, 2012, p. 83).

Fitzpatrick (2012) used the Lex30 cues to elicit a set of responses - four for each cue - from a single Chinese L1 participant during his eight-months study at a UK university. An identical version of the task was administered on six occasions at six-weekly intervals. At each timepoint in this longitudinal study the responses were treated as a vocabulary sample representative of the learner's productive lexicon. As explained in Chapters 2, 3 and 4, Lex30 uses 30 cue words which are selected from the first 1000 most common English words (Nation, 1984) possessing a number of

important criteria. These stimulus words trigger semantic concepts which Fitzpatrick maintains “will be reasonably consistent over time” (2012, p. 83) although the lexicalization of the concepts the learner wants to express may well vary considerably. Over time a learner’s ability to match these concepts with L2 responses will change as will the vocabulary and the degree of accuracy with which they can produce L2 words. The main aim of the study was to use an established investigative tool (Lex30) in a new way to try and trace stages in the “micro-development” (Churchill, 2008, p. 339) which might further reveal the complex processes taking place during lexical acquisition.

The WAT in this case was administered by computer and the participant provided up to 120 responses to the 30 cue words on each occasion. The six data sets were analyzed in order to investigate: (1) knowledge of written form (orthography), (2) knowledge of word parts (affixes) (3) knowledge of a word’s form and meaning and (4) knowledge of native-like associations. For a comparison with the experiment described in this chapter we are particularly interested about changes in orthography (part 1). Two important findings of the study are presented here.

- (1) The single subject produced 630 tokens over six test times (an average of 105 out of a possible 120 words per test) of which 88 tokens were misspelled (representing 13.97% out of 630 words). Fitzpatrick found that, on the whole, the subject’s spelling did not appear to improve or decline over the test period.
- (2) Fitzpatrick draws our attention to the fact that cues used with a WAT will sometimes elicit the same response at successive test times, and that this can prove interesting for analysis. On 13 occasions words were spelled incorrectly during an early test and then subsequently spelled correctly during a later one. Examples of items that were initially spelled incorrectly, but then stabilised into correct spelling, were: *rubbish*, *vegetable*, *relax*, *vocabulary*. However, the opposite case was also evident with 10 cases where words were first spelled correctly and then underwent a decline in spelling accuracy (Fitzpatrick, 2012, p. 87). Fitzpatrick noted that such spelling inconsistencies might perhaps be explained by variations in learner interlanguage where the types of errors in language produced by learners can change over time. Research by Romaine

(2003) seems to suggest that learners are liable to switch between a range of correct and incorrect forms over lengthy periods.

According to Ramanan (2016) this type of variability seems to be most common among language beginners, and may be entirely absent among the more proficient learners while Ellis (1999) suggests that “because acquisition entails item learning, free variation necessarily occurs before learners bootstrap to a system from the items they have learned” (p. 476). Fitzpatrick hypothesizes that many of the unusual or unlikely orthographic forms her subject produced was because he had not yet learned to bootstrap the spelling of unfamiliar words onto letter combinations of words that he already had knowledge of.

The author concludes that her study shows that a WAT, repeated at equal intervals in a longitudinal task, can be a useful tool for identifying and focusing on stages in the micro-development of the learners’ lexicon. It is apparent that in reality the acquisition process is non-linear and somewhat chaotic. As with many single case studies, there might be a level of concern about extrapolating from limited data to make generalizations about the learning process. Designing a larger experiment with multiple subjects might go some way towards satisfying this.

6.2.3 Cook (1997) and the ‘Dual-route’ model

In the previous section we have seen that orthographic changes can occur over time but an important question is the process by which a word’s spelling is acquired in the first place. What kind of influence does the English L2 learners’ first language have on writing and spelling ability? Cook (1997) conducted research into L2 users and English spelling and describes using a “dual-route” model to explain how new words are acquired. Such a model, he suggested, might be useful for describing the relationship between the sounds and spelling of different languages. The phonological route that this model describes is the relation of individual written *letters* to sounds. A number of rules usually exist for these letter to sound connections and the system is used as a default by the learner when new words are first encountered in a language. The visual route, on the other hand, is explained as one where whole *words* are accessed at a single time without reliance on phonology. The pronunciation of words like “yacht” and “cough”, for example, are learned by visual means rather than attempting one letter or syllable at a time. Cook maintains

that the more frequently used words in a learner's lexicon become then the more likely they can be accessed through such visual means (1997, p. 478).

Differences between L2 language systems can be also distinguished by using this dual-route model. Character-based scripts such as Chinese use a visual connection between the observed character and its meaning. Alphabet-based language scripts like Italian, Finnish, Spanish and to a lesser degree, French, have forms of pronunciation which are more predictable. The dual-route system might also provide an explanation for the differences in reading acquisition rates as well the ability to spell in different languages. Research has shown that languages which strongly adhere to spelling-sound rules are easier for children to learn how to read as they contain fewer exception words (Sprengrer-Charolles, 2011). The Spanish language's adherence to phonological rules, for example, can account for the fact that Spanish-speaking children exhibit a higher level of performance in nonword (or pseudoword) reading, compared to English and French-speaking children. A number of researchers have remarked on the differences between other L1 groups. Bebout (1985), in a study involving Native Speaker (NS) children and Spanish ESL learners, noted that Spanish speakers made more errors involving consonant doubling, whereas the native English speakers made more errors involving the unstressed vowel schwa (/ /) and the grapheme silent /e/. He argued that these differences stem from the language backgrounds and resulting spelling strategies of the 2 groups. Likewise Cook (1999) found that French speakers wrongly double consonants (e.g. *coming*) and substitute vowels (e.g. *materiel*) while common Chinese mistakes are the omission of consonants as in *subjet* and the addition of <e> as in *boyes*. Although the literature concerning how L1 characteristics can affect L2 vocabulary acquisition and spelling accuracy is still limited there is evidence to suggests that it may be an important factor. With a language like Japanese, which relies on both phonetic (kana) and visual (characters or kanji) as routes towards understanding the pictures, it seems more confusing. This will be explored in more detail in the following section.

6.2.4 Japanese L1 speakers and English orthography

English users usually employ a combination of methods when processing words with more frequently occurring words accessed visually but with seldom encountered words processed phonologically. Japanese users access Chinese-style

characters visually but take a different approach with syllable-based scripts or kana (hiragana and katakana) where a phonological route is followed to gain understanding. There is now a realization that the two routes are not distinct and that using a single dimension of orthographic depth may help provide a clearer explanation (Cook, 1997, p. 480). Some languages can be identified as being orthographically deep, such as Chinese, and tend to be meaning-based. Others are more sound-based, like Italian and Spanish, and can be described as orthographically shallow.

An important question is how L2 learners' knowledge of a language's phonological (letter to sound) rules and the processing of individual visual items is affected by their own L1 systems of spelling and pronunciation. L1 speakers of an orthographically deep language like Chinese might expect to have difficulty following a phonological route in more orthographically shallow language. Chikamatsu (1996) demonstrated that Chinese L1 speakers relied on visual strategies when learning the Japanese syllabic kana in contrast with English L1 speakers who relied on phonological strategies. The conclusion seems to be that L1 speakers of orthographically deep languages might find it easier to acquire languages through visual rather than phonological means.

It is uncertain which particular strategy Japanese L1 speakers might employ when learning English. Proper understanding of written Chinese-style characters demands a high degree of visual processing but understanding syllabic kana relies more on phonetic understanding. Bearing this in mind it would be interesting to see if Japanese L1 speakers display any unique patterns of learning development and language acquisition, particularly with writing and spelling, which are clearly different to those from other L1 groups. There are two studies that might be relevant here. I have discussed Cook (1997) earlier (see section 6.2.3), who looked at the spelling of adult L2 users of English and compared it with native L1 users, both children and adults, using data from learner corpora and EFL tests. Comparisons showed similar error rates in L1 children and L2 adults. The categories of errors included letter insertion, omission, substitution and transposition. Cook also examined the distribution of errors across certain L2 user groups and found that Japanese L1 speakers tended to confuse <l> and <r> in words and added vowels such as <e> or <a> at the end of nouns. He concluded that many of the errors reflected problems with sound / letter correspondences.

More recently, Gunion (2012) examined 15 Japanese students studying at a university in the UK for periods lasting between 6 months and 3 years. A 53-word spelling test (adapted from Okada, 1999) was administered and an error categorization system (adapted from Cook, 1999) revealed a pattern of misspellings perhaps due to the participants' Japanese L1. The patterns found were similar to Cook's and included < l > and < r > substitutions and vowel insertions. The well-known Japanese difficulty of being able to distinguish between the pronunciation of < l > and < r > (Cochrane, 1980) was probably the most commonly occurring feature in both studies. As well as using characters and kana, the Japanese language also uses *romanji* which is a system for writing both styles in Roman script. Romanji is used in a context where Japanese text is targeted at non-Japanese speakers and is also used to input Japanese words into word processors and computers. Many < l > and < r > words have conventional romanji spellings of imported loan words which are different from English, for example:

<i>Kana</i>	<i>romanji</i>	<i>English</i>
サラリー	sararii	salary
カレンダー	karendaa	calendar
ガラス	garasu	glass

The explanation for Japanese learners' difficulty with spelling < l > and < r > words may be partly due to alternative writing systems rather than simply pronunciation difficulties. There are two further complications that may occur. Firstly, it has been estimated that 18% of the Japanese language consists of loanwords from languages other than Japanese or Chinese including words of mixed origin and the made-in-Japan pseudo-English known as *wasei eigo* (Ebied, 2016). Learners on some occasions might mistakenly assume that a particular loan word comes from English when in fact it originates from another language. A common example of this is the word is the French word *pain* meaning bread. It is written in kana as パン and Romanized as: *pa-n*. Another is the imported Portuguese word *salada* meaning salad. It is written in kana as : サラダ and romanized as: sa-ra-da. Use of such words can often lead to misspelling by L1 Japanese users as they are wrongly assumed to originate from English. The second problem is that there are several systems for transliterating or romanizing Japanese into the Latin alphabet

(Smith & Schmidt, 1996). *Nippon-shiki* has become more predominant in recent years not least because of support from the Japanese Ministry of Education and Science (MEXT). *Nippon-shiki* is considered the most regular of the romanization systems because it maintains a strict “one kana to two English letters” form, is based on Japanese phonology and is thought most suitable for incorporating loan words into Japanese. *Hepburn-shiki* is based on English phonology and is effective for transliterating words from Japanese into English. *Hepburn-shiki* uses consonants that approximate those used in English. English or romance language speakers will generally be more accurate in pronouncing unfamiliar Japanese sounds with *Hepburn-shiki* than *Nippon-shiki* (Kindaichi et al., 1988). The main question about the use of these alternative systems of romanization is the effect that it may have on different Japanese L1 learners and their abilities to write and spell English words. An area for future research might look at younger Japanese L1 learners (or learners more familiar with *Nippon-shiki*) and see if they display similar patterns of L2 language development to those learners who may be more familiar with the older *Hepburn-shiki* kana style.

For Japanese L1 EFL learners, attempting to spell words with sounds that do not exist in their native language, is likely to be a serious challenge. To the author’s knowledge most studies concerning speaking and spelling to date have looked at the issue from the viewpoint of the influence of spelling on speaking and pronunciation skills (for example, Basetti et al., 2015 and Hayes-Harb, 2021). Instead, it might be worthwhile considering the opposite view and investigating whether learning accurate pronunciation can help improve spelling or even if other factors such as word familiarity or memorization are important. Tuladhar and Akatsuka (2017) examined this very point, basing their investigation on three wordlists comprising of 10 carefully selected categories of commonly misspelled words (shown in Table 6.1). The results of their study might inform us if any gains in speaking ability might also benefit learners’ ability to spell. Overall, the study tested the students’ ability to spell words after hearing and repeating their pronunciation with 100 Japanese L1 university students taking part. The three wordlists can be seen in Table 6.1.

Firstly, the words on list ‘A + 1’ were written down by all students after each word was repeatedly pronounced 3 times by the teacher. Soon after, students were divided into 2 groups. The first group of 50 students was given word list ‘B’ and the second group of 50 students was given word list ‘C’ (Table 8.2). Students were then

asked to pronounce each word as they normally would. They pronounced the words in their respective lists 3 times so that students of the opposite group could write the spelling.

Table 6.1

Categories of commonly misspelled words. After Tuladhar and Akatsuka (2017, p.99)

No. of words	Category	A + A1	B	C
1	borrowed words	Rest <u>au</u> rant	Hand <u>ker</u> chief	Al <u>co</u> hol
2	m [m], n [n], r [r], l [l]	Cont <u>em</u> por <u>ar</u> y	Com <u>pl</u> ain	Com <u>fo</u> rtable
3	ar/ir [ɑə-], [ə-]	F <u>ir</u> m	F <u>ar</u> m	H <u>ar</u> m
4	cial/ -tial [ʃ]	R <u>aci</u> al	S <u>oci</u> al	M <u>arti</u> al arts
5	or/ -er [ə]	Doct <u>or</u>	Invent <u>or</u>	Conduct <u>or</u>
6	kn (silent letter)	<u>Kn</u> it	<u>Kn</u> owledge	<u>Kn</u> ock
7	ch [tʃ]	<u>Ch</u> urch	<u>Ch</u> oice	<u>Ch</u> ocolate
8	double letter	Comm <u>un</u> ication	Emplo <u>yy</u> ee	Traff <u>ic</u>
9	memory	Pron <u>un</u> ciation	Let <u>tu</u> ce	Wed <u>ne</u> sday
10	ph [f]	<u>Ph</u> ysics	<u>Ph</u> ysical	<u>Ph</u> ilosophy

One week later all students re-wrote the words on list ‘A + 1’ while carefully listening to the teacher pronouncing them. This time, as well as listening to the teacher pronouncing them, they also wrote the words down after pronouncing them to themselves carefully three times. This final list of words that the students wrote down can be seen in Table 6.1.

There were several important findings. Firstly, Tuladhar and Akatsuka (2017) discovered that students were familiar with many of the word list items, but they could not spell them correctly because of the difficulty of maintaining a standard and stable pronunciation norm when using *katakana*. *Katakana* is the Japanese syllabic script used for creating written expressions of words borrowed from other languages. What this means is that on some occasions they were not easily able to recognize or identify the word. Secondly, the pronunciation by the teacher, the fellow student and by the speller themselves proved that accurate pronunciation can be useful in remembering the spelling of certain categories of words but not all categories. For example when words in list ‘A + 1’ were rewritten after pronouncing the word after

the teacher, the results showed that, though this had a positive effect on all categories of words, doing this activity had less impact on some categories than others. For example it had little effect on category 1 (borrowed words) perhaps because Japanese L1 EFL learners sometimes tend to spell English words the way they would write in katakana. According to Tuladhar and Akatsuka, one of the limitations to using katakana script is that it fails to express certain English half sounds as in “taxi”, “school” or “floor.” On the other hand, exposure to correctly pronounced words benefited other categories with regards to spelling. For example, category 6 (*kn* – silent letter), category 7 (*ch* - sound [tʃ]) and category 10 (*ph* - sound [f]).

6.3 Towards an orthographic experiment

This chapter reports an experiment using the Lex30 task to elicit a vocabulary sample at a range of time points before, during and after a short-term SA programme. With a L2 learner group of various proficiency levels but similar in age and background, it will attempt to observe, in a similar manner to Fitzpatrick (2012), some of the orthographic developments that can take place over time. Following Zareva (2012) it will also try to discover if the number and proportion of a learner’s partially-known vocabulary items is dependent on their proficiency level. Finally, it will comment on the patterns of English spelling to see if they are indicative of Japanese L1 learner behaviour.

6.4 Research questions

The study reported in this chapter will attempt to answer the following questions:

- (1) Do SA learners’ spellings tend to improve, decline or remain the same over the course of the SA programme? In particular, is this the case where the same response is given to a cue at multiple test times?
- (2) If making the assumption that misspelled vocabulary items indicates some degree of partial vocabulary knowledge while correctly spelled items indicate a more complete orthographic knowledge is it then the case that the number or proportion of partially-known words a learner knows is dependent on their proficiency level? Are lower L2 proficiency subjects likely to produce more

misspelled words in proportion to the total number of words produced than higher L2 proficiency subjects?

- (3) What is the nature of the pattern of spelling errors that subjects make and what might this tell us about the impact of their L1?

6.5 Methodology

6.5.1 Subjects

As previously described (in Chapter 4 and 5) 38 Japanese L1 EFL students travelled to the UK for a 15-day SA experience. The group was relatively homogenous with all participants aged 18-19, female and with proficiency levels ranging from advanced beginner learners (Common European Framework of Reference (CEFR) Upper A1 : TOEIC 230 – 280).

6.5.2 Task administration

A detailed report on the administration of the Lex30 productive vocabulary task has been given earlier in chapter 2 of this thesis. To briefly summarize, identical versions of the same Lex30 task (Meara & Fitzpatrick, 2000) which was used by Fitzpatrick (2012), was administered on four occasions before, during and after SA. The task was first administered 23-days before the start of the SA programme and then 1 day before departure for the UK. It was again administered for the third time 2 days after arrival back in Japan and then finally after 21 days (see section 4.3.3 and Table 4.1). Although the Lex30 task gives each participant the opportunity to produce 120 words, most produced considerably fewer than this (see Table 4.2).

6.5.3 Analysis and word set classification

In order to address research question 1 the analysis initially took a group perspective with all the words produced by each participant being added together at each time point. The following steps were taken. (1) The total number and percentage of misspelled words as well as correctly spelled words was noted to see if there is any discernable decline or improvement in spelling accuracy before, during or after the SA experience. (2) “Sets” of three or four similar, identical or near-identical words were identified. Each word set comprises of words which are

produced by the same subject in response to the identical cue word at different test times. Each word set is classified into one of 8 categories:

1. Four identical words are produced (by the same participant at each time point) and correctly spelled.

e.g. Cue word *attack*: (1) volleyball (2) volleyball (3) volleyball (4) volleyball

2. Three identical words produced (by the same participant at three out of four time points) and are correctly spelled.

e.g. Cue word *attack*: (1) enemy (2) enemy (3) enemy (4) _____

3. Four identical words (produced by one participant at each time point) are recognizable but incorrectly spelled.

e.g. Cue word *board*: (1) surfin (2) surfig (3) surffing (4) surfin

4. Three identical words (produced by one participant at three out of four time points) are recognizable but incorrectly spelled.

e.g. Cue word *cloth*: (1) sox (2) soccs (3) soks (4) _____

5. Signs of spelling improvement. The first one to three words are recognizable but incorrectly spelled. The last one to three words are correctly spelled.

e.g. Cue word *fruit*: (1) derishious (2) delicious (3) delicious (4) delicious

6. Signs of spelling improvement then decline. The first one or two words are recognizable but incorrectly spelled. The next one or two words are correctly spelled. The last one or two words are recognizable but incorrectly spelled.

e.g. Cue word *potato*: (1) frid (2) fried (3) fried (4) freied

7. Signs of spelling decline then improvement. The first one or two words are correctly spelled. The next one or two words are incorrectly spelled. The last one or two words are correctly spelled.

e.g. Cue word *map*: (1) picture (2) pcture (3) picture (4) picture

8. Signs of spelling decline. The first one to three words are spelled correctly. The last one to three words are recognizable but incorrectly spelled.

e.g. Cue word *hope*: (1) peace (2) peace (3) peace (4) pease

6.5.4 Proficiency groups and partial knowledge

In order to address research question 2 the 38 subjects were divided into two separate groups on the basis of approximate language proficiency in the same way as described in the previous chapter (see section 5.4). To remind the reader this is due to the fact that none of the SA participants had taken any form of established proficiency test (such as TOEIC) at the beginning and end of their SA experience and therefore it was decided to use a more approximate means to measure language ability. The main object in dividing the group in this way is to establish if the number of words produced along at each time point along with misspellings might shed light on the proportions of partially-known vocabulary to known vocabulary and proficiency level.

6.6 Results

In this section the results of the study are reported, with each research question addressed in turn.

6.6.1 Changes in learners' spelling during SA (RQ1)

The 38 subjects produced a total of 8996 tokens over four time points (an average of 59.19 words per subject for each test). Of these, 850 tokens were misspelled (representing 9.45% out of 8996 words). Looking at Table 6.2 below it appears that the number of misspelled items neither improved nor declined over the test period which is similar to Fitzpatrick's (2012) findings although the subjects seemed to produce fewer misspelled word on average than in the single case study (9.45% of 8996 as compared to 13.97% of 630 in Fitzpatrick's study). The

proportion of misspelled words to total words produced at each point, however, seems to gradually decline from Time 1 to Time 3 before rising slightly again at Time 4 as can be seen in Table 6.3.

Table 6.2

Number of responses and misspelled words given at each test time

	Time 1	Time 2	Time 3	Time 4	Total
Total number of words	1730	1972	2705	2589	8996
No. of misspelled words	212	200	218	220	850

Table 6.3

Percentage of items misspelled at each test time

	Time 1	Time 2	Time 3	Time 4	Total
Percentage misspellings	12.3%	10.1%	8.1%	8.5%	9.4%

Table 6.4 below shows that 273 sets of four and 390 sets of three identical and correctly spelled words are produced in response to the same Lex30 cue while on far fewer occasions (17 and 26 times respectively) sets of three and four words appeared which were all misspelled. The way in which and the degree to which the words were misspelled varies considerably.

Table 6.4

Groups where identical words appeared (Groups 1 to 4)

All words correctly spelled		All words misspelled		Total
(1) 4 times	(2) 3 times	(3) 4 times	(4) 3 times	
273	390	17	26	706

Using the system of classification outlined in the methodology Table 6.5 looks at the number of word sets where there is a combination of correctly spelled and misspelled words.

Table 6.5

Groups with both correctly spelled and misspelled words (Groups 5 to 8)

(5) Spelling improved (See table 5 below)	(6) Spelling worsened before improving.	(7) Spelling improved then worsening	(8) Spelling worsened	Total
64	22	12	17	115

Table 6.6 below shows 64 examples of instances of spelling improvement across three or four time points. The original cue word is on the far left followed by the word that the participant was attempting to spell. The four (or in some cases three) attempts at spelling the attempted word are given with misspelled versions written in **red** and the correctly spelled ones in **black**.

Table 6.6

Spelling improvement examples

<i>Lex30 cue</i>	<i>Word attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Attack</i>	[volleyball]	volley ball	volly ball	volley ball	volleyball
	[action]	acsshon	action	action	
	[volleyball]	vollyball		volleyball	volleyball
	[rugby]		ragby	ragby	rugby
	[baseball]	basseball	baseball	baseball	baseball
<i>Board</i>	[cleaner]	creener	cleaner	cleaner	
	[blackboard]	brackboard	black board	blackboard	blackboard
	[black]		balack	black	black
	[black]	brack	brack	black	black

Table 6.6*Spelling improvement examples – continued*

<i>Lex30 cue</i>	<i>Word attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Close</i>	[shop]	shope		shop	shop
	[door]	doore	door	door	door
<i>Cloth</i>	[skirt]	skurt		skirt	skirt
	[skirt]	skart	skirt	skirt	skirt
<i>Cloth</i>	[skirt]	skart	skirt	skirt	skirt
	[socks]	soks	socks	socks	
	[shoes]	shouse		shoes	shoes
<i>Dig</i>	[world]	warld	world	world	
	[mountain]	mauntain	mountain		mountain
<i>Hope</i>	[dream]	dreem	dreem	dream	dream
<i>Dirty</i>	[toilet]		toillet	toilet	toilet
	[water]	warter	warter	water	water
<i>Disease</i>	[sick]	sic	sic	sick	
	[headache]		headach	headach	headache
<i>Fruit</i>	[strawberry]	starrowberry		strawberry	strawberry
	[orange]	orange		orange	orange
	[strawberry]	strewberry	strawberry	strawberry	strawberry
	[delicious]	derishious	delicious	delicious	delicious
	[pineapple]	pinapple	pinapple		pineapple
	[strawberry]		stroberry	strawberry	strawberry
	[strawberry]		starawberry	strawberry	strawberry
<i>Furniture</i>	[chair]		chaer	chaer	chair
<i>Habit</i>	[manner]		manar	manner	manner
	[wash]		washe	washe	wash
<i>Hold</i>	[seat]		seet	seat	seat

Table 6.6

Spelling improvement examples – continued

<i>Lex30 cue</i>	<i>Word attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
	[museum]		museumu	museum	museum
<i>Hope</i>	[dream]	dreem		dreem	dream
<i>Kick</i>	[football]		foot ball	football	football
<i>Map</i>	[subway]		subwey	subway	subway
<i>Map</i>	[useful]	usuful	usuful	useful	
	[road]	oad	road	road	
<i>Pot</i>	[water]		warter	water	water
<i>Potato</i>	[fried]	fride	fried	fried	fried
	[delicious]	derishious	delicious		delicious
	[carrot]		calotte	carrot	carrot
	[vegetable]		vetitable	vegetable	vegetable
<i>Rest</i>	[vacation]		bacance	vacance	vacation
<i>Rice</i>	[white]	whaite	white	white	white
	[delicious]	delishous	delicious	delicious	delicious
	[delicious]	derishious	delicious	delicious	delicious
	[white]	wite		white	white
<i>Science</i>	[technology]	technorogy	tecnology	tecnology	technology
	[dangerous]		dengerous	dangerous	dangerous
<i>Seat</i>	[priority]	printo	priority	priority	priority
<i>Spell</i>	[difficult]		dificult	difficult	difficult
	[difficult]		dificcult	difficult	difficult
<i>Television</i>	[drama]	doram	doram	dram	drama
	[drama]	dorama	drama	drama	drama
<i>Tooth</i>	[white]	wite	white		white
<i>Window</i>	[cloud]	claud	cloud	cloud	

6.6.2 The impact of proficiency level on spelling (RQ2)

Zareva (2012) found that her L2 intermediate group had a greater partial knowledge in proportion to complete knowledge of the stimulus words used in her study compared to the L2 advanced group. In this experiment the 38 subjects were divided into two groups (n=19 each) on the basis of their L2 proficiency level. As explained earlier this was decided according to the total number of infrequently occurring words each subject was capable of producing when undergoing the Lex30 tasks. It is evident from looking at Table 6.7 below that the lower proficiency L2 group produced a higher percentage of misspelled words compared to the higher proficiency group perhaps indicating, as Zareva found, that they could have a higher proportion of partial vocabulary knowledge in their overall lexicon.

Table 6.7

Percentage of misspelled words out of total words: High/Low proficiency

<i>High proficiency</i>	Time 1	Time 2	Time 3	Time 4	Total
Words produced	1086	1224	1658	1546	5514
Misspelled words	102	104	95	100	401
Percentage words misspelled	9.4%	8.5%	5.7%	6.5%	7.3%
<i>Low proficiency</i>	Time 1	Time 2	Time 3	Time 4	Total
Words produced	656	746	1052	1055	3509
Misspelled words	103	92	114	111	420
Percentage words misspelled.	15.7%	12.3%	10.8%	10.5%	12.0%

6.6.3 Patterns of spelling errors and the impact of L1 (RQ3)

Finally, mention must be made of the third research hypothesis which asks what the pattern of spelling errors revealed might tell us about the participants and their L1.

(1) *Substitution of consonants*

Judging from several errors made in the study it seems that Japanese speakers are uncertain whether to use < l > or < r > in many cases.

creener	-	cleaner	brack board	-	blackboard
derishious	-	delicious	calotte	-	carrot
technorogy	-	technology			

The most obvious explanation is the well-documented Japanese difficulty in distinguishing between the pronunciation of < l > and < r > (Cochrane, 1980). Cook (1997) provides one explanation why this might be the case. As well as using characters and kana, the Japanese language also uses *romanji* which is a system for writing both styles in Roman script. Romanji is used in a context where Japanese text is targeted at non-Japanese speakers and is also used to input Japanese words into word processors and computers. Many < l > and < r > words have conventional romanji spellings of imported loan words which are different from English and may be partly due to alternative writing systems rather than simply pronunciation difficulties.

(2) *Extra vowel insertions*

Where there is a consonated cluster such as < dr > or < str > there is a tendency to insert an extra vowel. This can be seen in the examples where an extra < o > is added to drama: *dorama* or an extra < a > to *starawberry*. Consonant clusters do not happen in Japanese and loan words from English typically have vowels inserted between consonants (e.g. Macdonalds becomes *Makudonarudo*).

(3) *Added vowels at the end of words*

Partly due to the influence of the way some kana are pronounced and partly due to the influence of imported loan words there are sometimes additional vowels added to the end of words. Most Japanese words end in a vowel and this likely affects the production of English words in the same way. There are several examples in Table 6.5 including:

<i>museumu</i>	-	<i>museum</i>	<i>shope</i>	-	<i>shop</i>
<i>washe</i>	-	<i>wash</i>	<i>doore</i>	-	<i>door</i>

6.7 Further discussion

Fitzpatrick (2012) achieved two important findings in her study. The first suggested that her experimental subject's spelling did not appear to improve or decline over the test period. Out of 630 tokens which were produced over the test period, 88 were misspelled (representing 13.97% out 630 words) but there was no discernable movement towards improvement or decline (the maximum number of misspelled at Time 3 was 19 while the minimum number was 3 at Time 2). In this experiment far more tokens (8996 in total) were produced by the 38 participants of which 850 were misspelled (representing 9.45% of words). The subjects seemed to produce fewer misspelled words on average than in the single case study although the proportion of misspelled words to total words produced at each point seems to gradually decline from Time 1 to Time 3 before rising slightly again at Time 4. This can be seen in Table 6.2. Although the length of the SA period was far shorter (15-days as opposed to 8-months) there does appear to be some slight improvement in average spelling ability. Explaining this in terms of simply new language being acquired is highly uncertain but perhaps it is more likely that some reactivation of dormant knowledge has occurred (Meara, 2005) or simply that further consolidation of previously learned items has taken place.

The second finding concerns the description of identical words which were elicited as a response at more than one time point. On a number of occasions words were spelled incorrectly during an early test but then subsequently spelled correctly during a later one. This form of spelling stabilization also occurred with this experiment on 64 occasions (see Table 6.4) although spelling declines were also evident where words which had been first spelled accurately were later misspelled (17 in all). Such spelling inconsistencies, as Romaine (2003) suggested, might occur as learners tend to switch between a range of correct and incorrect forms over lengthy periods of time. This indication of free variation in the use of language features such as spelling may show that many learners have not yet mastered complex rules with which to control accurate production. Fitzpatrick's observation that the reason that unusual orthographic forms are produced by some learners is that

they have not yet learned to bootstrap unfamiliar patterns of spelling onto letter combinations they are already familiar with might be an accurate assessment.

Comparisons with Zareva's (2012) findings are more difficult to make. There are perhaps two points to mention here. Firstly, her study found that the L2 intermediate group admitted that they were unfamiliar with most of the words they were given at the start (72.3%). Of the remaining words (27.7% of the total) which they perceived at least some partial knowledge of, it transpired after further testing they had some knowledge (either partial or full) of a very small proportion of these. This was in contrast with the advanced L2 and NS groups who said that they were perhaps unsurprisingly unfamiliar with a much smaller number of given words at the start (55.4% and 50.8% respectively) and that it turned out that they knew (either fully or partially) a much higher percentage of the remaining ones. From this it is possible to draw the conclusion that lower L2 proficiency learners perceive that they have a greater partial knowledge than other proficiency groups but in reality, after further clarification and testing, this may turn out not to be the case. The present experiment divided the 38 participants equally into groups of higher and lower proficiency L2 learners (see Table 6.6). The higher proficiency group misspelled a smaller percentage of the total number of words that they produced (an average of 7.3% over 4 time points) than the lower proficiency group (12.0%). This might indicate that like Zareva's intermediate L2 group the lower proficiency group in this experiment perceived that they had knowledge of a relatively high number of words but when it came to demonstrating that knowledge by accurately spelling those words they were less able to do so.

The second point to make about Zareva's (2012) study is that it attempted identify partially known vocabulary by (1) pre-selecting a series of medium and high frequency words that were supposedly representative of the lexicon and (2) applying a meaning clarification checklist to assess learners' true (and perceived) knowledge of these words. As mentioned earlier (see section 6.2.2) the study succeeded in investigating four out of the seven aspects of Richards' 1976 word knowledge framework including understanding a word's semantic value, having a knowledge of its syntactic properties, knowing which other words are likely to be associated with it and underlying word forms. The present experiment lacked the same series of graduated word knowledge clarification steps and in doing avoided covering many aspects of Richards' word knowledge framework (though see chapters 5, 7 and 8 of

this thesis for analyses that address other aspects of word knowledge). The missing aspects would almost certainly include knowledge of a word's syntactic properties and some understanding of word's semantic values. The advantage of using a tool like Lex30 is that it can spontaneously stimulate participants into producing a sizable sample of vocabulary with comparatively little effort. One concern about this, however, is that the degree to which many of the words that are included in this large sample are actually known, remains unclear. Spelling accuracy is certainly one important aspect of word knowledge but there are many others some of which will be pursued in the following chapters.

6.8 Conclusion

An interesting finding by Okada (1999) is that Japanese SA participants appear to make fewer spelling mistakes than their contemporaries studying in a home environment. He discovered that 71.5% of all 795 words collected in a specially designed test were spelled correctly by students who had been, or were participating in SA. This compared 55.6% of the same number of words with a sample collected in Japan. This example of research indicates that there does seem to be some improvement in spelling accuracy during the SA process. The experiment conducted in this chapter suggests a similar increase in spelling accuracy together with other forms of language improvement despite the very short time frame (15-days) of this particular SA experience. As explained in chapter 2, this may be partially due to students undergoing a form of skill reactivation which would help explain why they can experience such rapid improvement in their skills. This reawakening of dormant vocabulary knowledge (including spelling) is perhaps partly due to the influence of an L2 environment (Meara, 2005). With Japanese students studying in the UK, many of whom are travelling abroad for the first time, the effects of such an environment should not be underestimated. The exposure to a large amount and variety of target language input, is likely to have an impact on sensitivity to distinctions between phonemes, and on pronunciation. This would support students in learning to distinguish between < l >, < r >, < b > and < v > sounds and remembering to omit an extra vowel on the ending of words like salad (and not salada) and museum (and not museumu), and in turn is likely to have at least some beneficial influence on spelling accuracy.

The experiment shows that there are signs of a gradual improvement in spelling accuracy during the course of an SA programme in terms of the percentage of words misspelled but there is still much room for improvement. It would be interesting to conduct a similar longitudinal experiment over a longer duration (like Fitzpatrick, 2012) to see if a more complete pattern of spelling development is revealed. Although it is helpful to trace the process of skill reactivation that may take place over the very short-term, with a longer-term study it might be possible to discover what real gains occur, and learn the whole story of how words are assimilated into the learner's lexicon. Tracking an individual word from its first appearance as a partially-known and misspelled fragment to a level where it is accurately used with confidence in a variety of contexts and meanings would represent a challenge but this might be achieved given enough time. Another point is that developing a simple method of vocabulary knowledge clarification (like Zareva, 2012) would go some way in improving our picture of the development of learners' level of knowledge of a word in the context of Richards' (1976) word knowledge framework.

Finally, it is worth mentioning that carrying out some form of categorization process of misspellings (e.g., Cook, 1997; Gunion, 2012; Okada, 1999) would be a useful exercise not least for pedagogical purposes. This might involve identifying the most common areas of misspellings and concentrating on particular rules, for example, the change of <y> to <ie> before <s> *carries*, so that troublesome groups of words or word-endings can be tackled. In the next chapter we will continue to explore the developing lexicon, as evidenced in Lex30 data, by investigating the semantic domains that seem most affected by the SA experience.

Chapter 7: Wmatrix and changes during SA

In this chapter I will continue to analyse the Study Abroad (SA) data produced by Japanese L1 EFL learners before and after their programme, this time using an innovative online tool known as Wmatrix. This tool is capable of comparing target corpora, in this case words produced by participants at time points before and after an SA experience, to see among other things, if there are any changes to the semantic domain that words belong to.

I will start by explaining some of the features of Wmatrix and give some examples of the way this tool has been used across a variety of fields to answer diverse questions. Then I will narrow this down to discuss some Wmatrix studies related to second language learning and see if they have found differences in language produced by both L1 and L2 language speakers as well as between cultural groups. I will evaluate Wmatrix's ability to trace some important changes including whether there are any differences in the number of words belonging to particular semantic domains. Will SA participants, for example, tend to produce more travel or sightseeing related vocabulary as a result of their experience?

Finally, the SA data will be subdivided into words produced at specific time points to see if any significant changes in the frequency of occurrence of individual words, Part Of Speech (POS) groups and semantic domains, occur over the course of a short-term SA experience. Where evidence of such changes exist an attempt will be made to compare them with the findings of earlier studies (Lin, 2014; Lin, 2017) and to explain them within the framework of a typical SA programme. This study will allow us to test our prediction that the application of Wmatrix to SA data may help us to identify certain specific characteristics in the kind of language SA participants produce and how these might change over time. We will investigate whether an SA experience can boost the acquisition of vocabulary knowledge from certain domains. Explaining these changes in terms of the linguistic and cultural influences that a particular L2 environment is likely to offer may contribute towards a better understanding of the design of future SA programmes.

7.1 Using Wmatrix

Wmatrix is a web-based software tool that can be used for corpus analysis and comparison (Rayson, 2008; see also <http://ucrel.lancs.ac.uk/Wmatrix>). Its features include the ability to analyze texts for Part Of Speech (POS) and semantic information and it can provide a useful web interface for the English CLAWS and USAS corpus annotation tools. It also incorporates standard corpus linguistic methodologies such as frequency lists and concordances. CLAWS (Constituent Likelihood Automatic Word-tagging System) is a tagger for POS annotation and can automatically assign POS fields respectively to each word in a corpus. Rayson (2008) reports that the accuracy rates of POS tagging are approximately 96 to 97% which can give us some confidence in its accuracy. USAS (UCREL⁵ Semantic Analysis System) is a tagger for semantic annotation and can automatically assign semantic fields. Wmatrix uses 21 major semantic categories and many minor sub-categories (in fact, 232 sub-categories according to Archer et al., 2003).

The major semantic categories are shown in Table 7.1 which is reproduced from Archer et al. (2003, p.2) and the complete tagset (reproduced from: <http://ucrel.lancs.ac.uk/usas/>) is shown in Appendix 11). The contents of categories are not always mutually exclusive and it is possible for words to fall into more than one of them (Archer et al., 2003, p.2). For example the word *bank* could belong to H (architecture) with a physical building or *bank*, I (money and commerce) with money belonging to a *bank* or W (world and environment) where the word could relate to the *bank* of a river. (Rayson et al., 2004, P. 2) The semantic concepts used for the categories are based upon general meaning, rather than on any psychological interpretation. Wmatrix's semantic categories were derived from the 1981 Longman Lexicon of Contemporary English (Archer et al., 2003, p.2), and although Wmatrix has continued to evolve since its original conception, the basis for its categories has not changed.

One feature of Wmatrix is that it can assist us with "Key word analysis". This may prove useful as it can help us identify individual words and semantic categories connected with language that are used during an SA experience. Key word analysis has been widely used in many areas of applied linguistics research, particularly with regard to the identification of language variation, genre and discourse

⁵ University Centre for Computer Corpus Research on Language (Lancaster University, UK).

Table 7.1*Wmatrix Major Categories. (reproduced from Archer et al. 2003, p.2)*

A general and abstract terms	B the body and the individual	C arts and crafts	E emotion
F food and farming	G government and public	H architecture, housing and the home	I money and commerce in industry
K entertainment, sports and games	L life and living things	M movement, location, travel and transport	N numbers and measurement
O substances, materials, objects and equipment	P education	Q language and communication	S social actions, states and processes
T time	W world and environment	X psychological actions, states and processes	Y science and technology
Z names and grammar			

characterization (Harvey, 2013; Lin, 2013; Scott, 2010; Scott & Tribble, 2006). Key words are those items that occur unusually frequently (positive key words) or unusually infrequently (negative key words) in a target corpus as compared with some norm (Scott, 2010). They are identified on the basis of statistical comparisons of words in a text or collection of texts against a reference set of words. In this way, any word that is found to be exceptional in its frequency in the comparison is considered key. One example of the use of keywords is given by Harvey (2012) who used a corpus of adolescent online health communication which he then compared with a general English corpus. Using identified key words, he examines how adolescents communicate the psychological distress of depression to health professionals. Through his research he managed to gain a deeper insight into young people's health problems and the linguistic particularities of online health communication.

Until recently key word analysis has only focused on lexical differences, rather than semantic, grammatical or functional differences. Baker (2006) criticises

this limitation pointing out that it is sometimes the case that certain words do not occur often enough to make a sufficient impact, and may be overlooked when compiling key word lists. Extending the key word method to also include information about POS and semantic domain levels, however, could help us extend the range of language coverage and provide further valuable information on particular language use.

Calculation of statistical significance for “keyness” is usually carried out using the Chi-square and Log-likelihood (LL) methods (Dunning, 1993). These both compare the differences between the observed values and the expected values: the greater the difference between the two values, the more likely it is that the relationship between the two items is not due to chance, but that other factors influence their relationship (McEnery et al., 2006). In this way, the items that have unusual characteristics in a target corpus can be identified by applying the Chi-square and LL tests. However, McEnery et al.’s (2006) study supports the use of the LL test as it does not assume that data is normally distributed. Dunning (1993, p.65) notes that “using the normal distribution overestimates the significance”; consequently, the use of likelihood (LL) ratios leads to very much improved statistical results, particularly when analyzing small volumes of text.

Returning to the research carried out in this thesis Wmatrix might help us scrutinise the language produced by SA participants, particularly in a UK setting. Wmatrix provides access to both the British National Corpus (BNC) written and BNC spoken sub-corpora frequency lists which can be used as reference corpora for comparisons with user-supplied corpora. Since it relies heavily on the BNC, and is U.K. based, Wmatrix employs British English spellings. This is highly relevant for the data collected for the experiments reported in this thesis. The tagger operates on multi-word units as well as single words, so it can deal with phrasal concepts (such as ‘more and more’). The results of keyword comparisons are produced in descending order of Log Likelihood (LL).

7.2 Interpreting Wmatrix’s output

Wmatrix has been applied to numerous research questions across a wide range of fields since its conception. In this next section we will look at four examples illustrating how the tool can identify certain specific characteristics in language. These include an analysis showing which linguistic and rhetorical techniques are

most typical of tourism promotion, a study into how two different political parties operationalize emotion in their discourse, a forensic study of suicide notes comparing topics used in real and fabricated examples and finally a study of politicians' discourse suggesting that the issue of climate change is amenable to straightforward policy action.

In the first example, Bianchi (2017) uses Wmatrix to show the importance of social media marketing to travel companies and how certain linguistic and rhetorical techniques are used to attract potential customers. In her study, Bianchi builds on the work of previous researchers which found that the language of tourism features particular characteristics (Dann, 1996; Gotti, 2006). Bianchi created a specialized corpus (FB Tourism Corpus) using Facebook posts produced by three tour operators based in the UK, the USA and New Zealand and compared this with a reference corpus which included examples of countries' promotional webpages and online travel agent advertising material. Using Wmatrix to identify a number of positive key words and also following Baker's (2006) call to extend the key word method to POS and semantic domain levels, she shows that the language used in Facebook posts differs in many respects from standard tourism promotional language in the use of pronouns, adjectives, imperative verbs and questions. Bianchi's experiment helps demonstrate that Wmatrix is capable of detecting the development of conversational strategies that tour operators use to convince potential customers that they have a knowledge of their personal needs and desires (Gretzel & Yoo, 2014).

Breeze (2019) uses Wmatrix for quite a different purpose by examining the use of emotive expressions in the language of official press releases provided by two UK political parties. Her study compares the frequency of specific semantic subcategories and focuses on the political visions of a future society. Building on work by Moffit (2016) and Van Leeuwen (2014) who identified the specific language elements that were more likely to provoke emotional responses, Breeze (2019) demonstrates the fact that certain politicians, who use such elements, are also often highly successful - particularly in the age of Youtube, short sound bites and Tweets (Frame & Brachotte, 2016). Unlike Bianchi's (2017) study, no reference corpora are used. Instead Breeze creates a number of corpora for a comparative investigation which are analysed by Wmatrix and semantically tagged using the USAS annotation tool (described in section 7.1). The results show that one of the political parties in her study has a higher total incidence of emotion-related words

(tagged 'E' in the USAS Semantic Tagset - see Appendix 11) and especially words classed as E3- (denoting *violent* or *angry*) and E4.1+ (*happy*). Breeze (2019) clearly identifies how specific emotional areas such as fear, anger and anxiety are invoked by the two parties.

The next example comes from the field of forensic linguistics. Distinguishing real from fabricated suicide notes is a challenging but critical element of criminal investigation and linguistic analysis using Wmatrix can help to inform this. Shapiro (2011) investigates the language of suicide notes comparing topics used in both versions. In particular, she is interested in seeing if there are any differences in lexical content, key words and semantic categories. Her findings indicate that the presence of certain content can distinguish suicide notes from other texts. Key words include such examples as *grieve, sad, upset, unhappiness, suffering, crying, misery, depressed, unhappy, pity* and *regret*. When looking at the UCAS tagset (see Appendix 11), it is clear that several categories are strongly represented such Z8 (*pronouns*), E2 (*like*), Z6 (*negative*), T1.1.3 (*time: general:future*), 'X2.2 (*knowledge*) and S4 (*kin*). Additionally, genuine suicide notes seem to be distinguished by their greater than normal use of pronouns, people's names and misspellings, indications of author identity such as age, gender and relationship with the intended reader and the use of terms of affection.

Finally, Wmatrix may also allow us to examine the differences between politicians' and scientists' presentation of certain complex issues. Evidence from corpus analysis suggests that politicians often present their arguments on the issue of climate change using mainly scientific and economic frameworks and rarely consider environmental or social factors (Fielding et al., 2012; Partington, 2012). Using corpus analysis together with Wmatrix, Willis (2017) investigates how UK politicians conceptualize climate change as a political issue, by creating a specialist Climate Change Bill (CCB) corpus and comparing this with reference corpora including samples of spoken political debate. Using Wmatrix, she compiles lists of under and overused keywords and groups words from each corpus into semantic fields allowing her identify trends and patterns. Her findings indicate politicians' clear tendency to present climate change as a technical issue, by using strongly scientific, technical and economic language with little consideration of social or environment issues. Willis also discovers that politicians' use of science is also

highly selective, with little discussion of abrupt or irreversible impacts, in contrast to prevailing scientific consensus.

What can we learn about Wmatrix from these studies? They might go some way to supporting its use in analyzing the language produced before and after an SA programme particularly in terms of whether the experience can have an impact on the language produced by participants in terms of keywords, the semantic domain to which the keywords belong and their POS category. Instead of producing words connected with areas such as tourist advertising or climate change, for example, Wmatrix might show if SA participants can demonstrate an increase in the knowledge of words used in specific travel situations such as shopping, restaurants or language likely to be used in conversations with a host family.

7.3 Wmatrix, corpus analysis and EFL

So far we have looked at studies which demonstrate Wmatrix's ability to identify specific language characteristics in variety of situations but which lie outside the area of EFL. Two of the studies, Bianchi (2017) and Willis (2017), use reference corpora consisting of online travel agent advertising material and samples of spoken political debate for comparative purposes. The other studies by Breeze (2019) and Shapiro (2011), make comparisons between the target corpora that they investigate. The latter two studies show that it may not always be possible to use appropriate reference corpora particularly where the subject material maybe unusually specialized, as in the case of suicide notes. However, using a carefully chosen comparison corpus may go some way towards eliminating interference caused by variables which may not make up the focus of investigation.

After examining Rayson's site which lists cases where Wmatrix has been used in many different kinds of research (<https://ucrel.lancs.ac.uk/wmatrix/#apps>) and also conducting an extensive online search, there appear to be few examples where this analysis tool has been used in EFL research. Many of the examples that Rayson gives are related to forms of discourse analysis relating to language used in topical areas such as gender, literature, business, politics and the environment. My own investigations have revealed only two instances where Wmatrix has been used which perhaps might prove useful for the experiment described in this chapter. Studies by Lin (2014; 2017) using Wmatrix and corpus analysis have sought to bring the use of this analysis tool into the area of EFL research. Lin (2014) looks at

samples of spoken discourse between groups of British and Taiwanese students, with the aim of exploring if there are any statistically significant differences in the use of grammatical categories. His second study (2017) builds on this by exploring the patterns of lexical features in both online and face-to-face spoken communication between members of the same cultural groups. In the next section we will look at both studies to see if they can inform my own research into the changes which may occur with Japanese L1 EFL learners during an SA experience. An initial glance shows that there are likely to be some differences: Lin, in both of his studies, carries out cross-sectional analyses with two quite separate cultural groups whereas our own experiment involves a longitudinal study with data gathered at different times from individuals within a single group. It does seem likely, however, since both involve non-native speaker participants, that there might be some similarities in language characteristics and cultural reflection produced by the participants involved. I will first present Lin's (2014; 2017) studies then in section 7.3.3 consider how it relates to my own.

7.3.1 Analyzing grammatical categories: Lin (2014)

Lin (2014) sets out to conduct key word analysis to identify features of corpora and to investigate whether there are any differences in the use of grammatical categories used between 40 Taiwanese EFL learners and 40 British secondary school students. As mentioned earlier in section 7.1, key words appear unusually frequently (or infrequently) in a target corpus as normally compared with a more general reference corpus. Lin proposes a "data-driven" approach which allows "macroscopic analysis" (performed on whole bodies of corpora) to inform microscopic analysis (concentrating on single linguistic features) (Rayson, 2008, p. 39). In this case the linguistic feature chosen for study is decided by the information gained from the data itself after it has been gathered. In this way specific linguistic features are first highlighted for further investigation and are then further informed by macroscopic analysis. In this particular study, the initial macroscopic analysis looks at differences in the usage frequencies of lexical and grammatical features in two groups of discourse involving the POS components of Taiwanese learners and native English speakers.

Lin (2014), referring to work carried out by Baker (2006) (see section 7.2), argues that research should not be limited to identifying differences in key words

contained within different corpora but should also extend further to identifying differences in their POS categories. He proposes that an investigation into the differences in POS categories may provide new information on the characteristics of particular language use. An early example of such an analysis was in a study carried out by Ooi et al. (2007) which examined Singaporean English weblogs and compared them with a large English corpus. Through Wmatrix-extended POS analysis they gained a greater understanding of cultural identities and gender amongst other variables.

In using a similar approach, Lin (2014) aims to see how EFL learners and native speakers each use their lexical resources. He uses a specialized corpus derived from a cultural exchange project, the British and Taiwanese Teenage Intercultural Communication Corpus (BATTICC), for his research. Participants, aged from 13 to 15 years old, carried out a series of face-to-face spoken and weblog written interactions in both Britain and Taiwan over a three-week period. BATTICC represents a particular genre of intercultural communication where participants are not given any specific guidelines regarding what to talk about or how to structure their interaction allowing them the freedom to create conversational topics of common interests where the main aim is to sustain communication.

Lin's methodology involved dividing BATTICC into separate Taiwanese and British participants' datasets and then using Wmatrix to conduct key word analysis to make statistical comparisons of the frequency lists derived from both target corpora. Then, by using the most frequently occurring key words, the corresponding POS category to which they belong could be identified. Wmatrix is capable of sorting key words into 140 separate POS categories (see Appendix 12) and, according Rayson (2008) can consistently achieve 96-97% accuracy (the precise degree of accuracy varying according to the type of text analyzed). In this way Lin was able to use Wmatrix to generate a number of key POS domains to identify if any grammatical categories used by Taiwanese participants are significantly different to those used by British ones.

The results highlight a number of grammatical differences between the two data sets (see Appendix 13.1), particularly the fact that many Taiwanese participants tended to use verb tenses incorrectly, for example *is* and *was*, and there was often mis-selection of verb form where tenses were not properly used. Although Taiwanese learners seem to have had difficulties in selecting the appropriate verb

tenses in English, it also appears that the level of written English that they produced is greater than their spoken and that they were able to use verb tenses more accurately. Another difference that POS analysis revealed was that interjections, which were usually used to fulfill a number of discourse functions, were used by Taiwanese participants with unusual frequency. Finally, ‘ditto tags’, a category of general adverb, were much more common with British participants and seemed difficult for Taiwanese students. “Ditto tags encode the notion that a token is not an individual unit, but rather is a (somewhat non-compositional) part of a larger ‘idiom’” (Dickinson, 2005, p.57). They comprised 0.53% (63 tokens) of the total number of tokens produced including expressions like *sort of* (19 tokens), *a lot* (14), *a bit* (12), *as well* (7) and *kind of* (3). Meanwhile the Taiwanese participants only produced 0.06 % (or 4 tokens) out of their total number with ditto tags like *a lot* (2), *as well* (1) and *of course* (1).

Lin’s (2014) study reveals the extent to which grammatical categories can occur unusually frequently or unusually infrequently in EFL learners’ discourse when compared with the language used by native speakers of English. His research findings underline the fact that there is some pedagogical merit in key domain analysis as it can help inform EFL researchers, educators and materials developers in the design of courses emphasizing spoken interaction. There are two further points that can be made about Lin’s (2014) study. Firstly, like my own study reported later in this chapter, he does not refer to any reference corpora with which to make comparisons. However, he examines differences between corpora produced by his two study participant “types”, whereas my own study involves a within-subject comparison, comparing the language performance of the same group of participants over time, before and after a major intervention activity (SA). As far as I can discover while there are a number of examples where Wmatrix is used to measure differences between various language user groups there are none which see the tool being used to measure longitudinal changes within the same group. Secondly, with regard to the form of analysis used, Lin (2014) limits his use of Wmatrix to making a comparison of the different frequencies of certain grammatical categories which occur within two target corpora. However, Wmatrix has a number of additional features that could yield yet more valuable information about patterns of vocabulary use during SA. These might include assessing levels of keyness for individual predetermined lexical items or tracing changes to particular semantic domains.

7.3.2 Analyzing key words and semantic domains: Lin (2017)

An important difference between Lin (2017) and his earlier study is that he begins to emphasize the intercultural dimension of language education and language learners' ability to interact with others. Employing a model of intercultural competence proposed by Byram (1997) and Byram et al. (2002) and developed by Fantini (2000; 2012), he lists concrete curricular objectives for the EFL classroom including intercultural knowledge and the ability to relate to others. Lin (2017) builds on work by Liaw (2006), who established a web-based environment to help university students develop both linguistic and cultural competencies. Liaw (2006) analyzed web-forum entries of Taiwanese EFL students corresponding with English speaking partners in the US and found an increase in the length and complexity of their writing. Lin (2017) seeks to build on the success of Liaw's (2006) study, demonstrating that intercultural competency, developing attitudes of curiosity, openness and critical cultural awareness can increase given the right circumstances.

The aim of Lin's (2017) study is to analyze intercultural discourse in online and face-to-face settings by examining keywords and semantic domains to discover which topics may be of interest to young people. Participants in his study include 30 Taiwanese pre-intermediate level EFL learners and 30 British secondary school English native speakers. Participants had an extended period of interaction with each other on an online discussion board before finally experiencing face-to-face group meetings during a three-week SA period. The study was not longitudinal with changes in the same participants being measured over time, but rather a four-way one seeking to concentrate on the differences between two cultural groups and between reference corpora. All spoken and written discourse was gathered and two corpora produced. These were the British and Taiwanese Teenage Intercultural Communication Corpus Online (BATTICC-O) (31,916 words) for online discussion and BATTICC-F (34,089 words) for face-to-face interaction. The first stage of methodology that Lin employs is making a statistical comparison of frequency lists derived from the two target corpora noting that this can offer an "immediate snapshot of the characteristics of a particular language variety" (Harvey, 2013, p. 57). However, Lin is cautious about word frequency lists stating that they sometimes fail to identify important differences between texts supporting earlier evidence found by McCarthy and Handford (2004). He uses two reference corpora to circumvent this

concern: CANELC⁶ (e-language corpus) and CANCODE⁷ (an informal spoken English corpus). Lin compares frequency lists compiled from the two BATTIC corpora with his reference corpora. The second stage of methodology uses Wmatrix to automatically assign POS and semantic fields respectively to each word in the semantic categories were “key”.

Initial results are interesting. Four frequency lists (BATTICC-O, BATTICC-F, CANELC and CANCODE) are generated and a four-way keyword analysis employed to highlight their main characteristics. Using Wmatrix, all the target corpora words are ordered by ‘keyness’ value (where they occur unusually frequently or infrequently in relation to both the reference corpora). An example of the first 10 most frequently items in BATTICC-O and CANELC is given in Appendix 13.2) The personal pronoun ‘*I*’ and its possessive form ‘*my*’ are the most prominent positive keywords occurring more frequently in the BATTICC-O corpus than in the CANELC. Other interesting words are ‘*haha*’ ranked the 15th which represents the sound of laughter. ‘*haha*’ can be found in 19 instances and is used by both British and Taiwanese participants. Many words reveal specific themes that participants are mainly concerned about during their online discussion including school life and what they particularly liked. Words like ‘*favourite*,’ ‘*like*’, ‘*play*’, and ‘*food*’, country names like ‘*Taiwan*’ ‘*UK*’ and ‘*England*’ and words connected with school life, school events and festivals and friendship possibly appear the most when the participants meet face-to-face. The key word list seems to be a snapshot revealing predictable contact domains that are frequently talked about by the participants.

Key semantic domains are identified in both BATTICC corpora using Wmatrix. 37 separate USAS tags were significantly under or over used between BATTICC-O and CANELC data (please refer to Appendix 11 to see the complete USAS tagset). Results of this semantic domain analysis are shown in Appendix 13.3. Most notable is the category Z8 (denoting *pronouns*). The next most notable categories are P1 (*education*), F1 (*food*), E2+ (*like*), K1 (*entertainment generally*),

⁶ CANELC stands for Cambridge and Nottingham E language corpus. The sample used in the study includes 500,000 words of asynchronous online discourse from a variety of sources such as online discussion boards blogs tweets and emails. (Knight et al., 2014)

⁷CANCODE stands for Cambridge and Nottingham corpus of discourse in English, a 5 million-word corpus of mainly informal spoken English. The corpus was developed as a joint project between the University of Nottingham and Cambridge University press.

S3.1 (*personal relationships*) and K2 (*music and related activities*). These domains are overused in that they tend to be used more in adolescent online communication than online communication generally. On the other hand the category I1 (*money*) was far less talked about perhaps being of less concern to adolescents still reliant on their parents' support. Z99 (*unmatched*) was also common perhaps caused by the large number of computer acronyms and abbreviations.

With BATTICC-F and CANELC data, 102 USAS tags were significantly under or over used. Notable categories or domains were Z2 (*geographic names*), E2+ (*like*), P1 (*education*), A13.3 (*degree: boosters*), F1 (*food*), W3 (*geographical terms*), W4 (*weather*). Underused or far less talked about domains were Z6 (*negative*, A3+ (*existing*) and A7+ (*likely*). When looking at both data sets (BATTICC-O and BATTIC-F) P1 (*education*), F1 (*food*) and E2+ (*like*) occur with high frequency indicating that these three topics are popular with both face-to-face and online communication. Lin suggests that the reasons for this are that the Taiwanese and British participants enjoyed talking about their own cultures and contrasting the differences between them. Aspects of participants' cultural background are also evident when comparing key items within semantic domains. For example, when Taiwanese and British participants produce a similar high number of items which happen to belong the same domain such as F1 (*food*), P1 (*education*) or K5.1 (*sports*) the actual words themselves used by each group tend to be very different. In the F1 (*food*) domain, Taiwanese participants produce words such *rice*, *noodles* or *tofu* while British participants refer to *potatoes*, *hamburgers* and *pizza*. With P1 (*education*) the cultural differences are more revealing. Lin notes that in Taiwan it seems more common to talk about *tests*, *exams* and *study* whereas in Britain *lessons*, *colleges* and *tutor* appear more relevant. The choice of words within each semantic domain seems to demonstrate cultural and social differences with regard to the lexical choices of British and Taiwanese participants.

The investigation of adolescent intercultural discourse using Wmatrix and specialized corpora (BATTICC-O and BATTICC-F) show the semantic categories that are key in online and face-to-face communication. Some of these categories represent themes or topics that were specific to the participants involved. With online communication, pronouns, education, food, likes, entertainment and personal relationships are identified as key. With face-to-face communication geographic names, likes, education and food are seen as more important. When taking both

forms of communication together, Lin (2017) finds that education, food and likes can be identified as the key semantic domains. Words produced for each semantic category seem to display a great number of cultural and social differences in terms of lexical choices by the different groups of participants. The study also reveals a number of cultural and social differences in the use of words and that semantic categories can contribute to a better understanding of learner vocabulary. Lin (2017) concludes that his research can help EFL teachers develop their cultural awareness and intercultural communication skills so that they can encourage their learners to more closely observe culturally relevant linguistic features.

7.3.3 Comparing Lin (2014; 2017) with my present study

Lin's two studies inform my own research in some important ways. The first of these concerns the selection of suitable reference corpora as Lin (2017) manages with CANELC and CANCODE. Creating a corpus using words from the Lex30 task completed by Japanese L1 test subjects under 'At Home' (AH) study conditions rather than during Study Abroad (SA) might certainly be possible. By doing this we then might be able to conclude that any differences in keyness and semantic domain occurring between reference and target corpora could be at least partially due to the influence of participants' experiencing a unique SA environment. However, there is difficulty in creating such a corpus due to the tendency for test subjects to vary their responses to Lex30 over time and taking into account the likely differences in subject profile and proficiency level. In lieu of producing such a specialized corpus, using small 'sample' corpora (for example, the BNC writing sampler mini corpus) provided by Wmatrix as part of its suite of analytical tools, might still be possible.

A second consideration is that Lin's (2017) study use corpora comprising of spoken and written discourse spontaneously produced through natural human interaction. Such discourse is more likely to share similar characteristics with a number of established reference corpora at least regarding the frequency of occurrence of certain POS categories. The SA data used for our own analysis bears little relation to natural discourse but is rather a list of words produced in response to a series of stimuli. Corpora consisting of words obtained by such fundamentally different methods are likely to vary in a number of different ways.

The difficulty with selecting a reference corpus for the purposes of comparison calls for a solution which will still allow us to make a useful analysis of

our data. As discussed before (in chapter 4) our data consists of words collected at four separate time points during our SA programme. Data collected at the first two time points (around three weeks before departure and near the actual departure date itself) reflects participants' vocabulary knowledge before their SA experience. A corpus created using such data might be able to act as a useful reference with which data gathered at subsequent time points could be compared.

A final consideration concerns Lin's (2017) interest in the intercultural dimensions of language and the way that Wmatrix identifies certain words produced by his Taiwanese participants as reflecting their cultural background. In a similar way we can see if the Japanese participants involved in our own study tend to choose different words within the same semantic domain than would be generally expected from a native speaker population. To give some examples, they are more likely to produce culturally determined answers for semantic domains such as those connected with food (F1), education (P1) and sports (K5.1).

From our assessment of previous research where Wmatrix has been used and more particularly in the case of Lin (2014; 2017) where it has been used in way that may be more relevant to my own experiment suggests there are a number of questions that this analytical tool can be used to answer. These will be discussed in the next section.

7.4 Research questions

Our research questions will consider the language data collected from our SA participants via the Lex30 task to assess its characteristics and whether any change might occur during an SA experience. The SA corpus is created using individual words given in response to a series of stimuli rather than using samples of general discourse. This is likely to give rise to important differences between corpora created by each method, and raises questions about whether it is possible for reference corpora to be used in an appropriate manner. Our first question (1) concerns our overall SA corpus, that is all words produced by participants before and after an SA experience, and whether this might be comparable with a reference corpus (in this case the "BNC sampler written informal") incorporated into Wmatrix's suite of tools. The first (1) research question will ask:

1. a. Do certain words appear more frequently or infrequently in an SA corpus than in the BNC sampler written informal reference corpus?
- b. Are certain POS categories over or underrepresented when compared to existing corpora?
- c. Are certain semantic domains over or underrepresented?

The second question will deal with language change. More specifically, is the language that SA participants produce after their experience in any way different to that which they produce beforehand? Will there be any difference in language in terms of individual words, POS categories and semantic domain? The second question (2) will ask:

2. a. Do certain words appear more frequently or infrequently in the pre-SA corpus than in the post-SA corpus?
- b. Are certain POS categories over or underrepresented in the pre-SA corpus compared to the post-SA corpus?
- c. Are certain semantic domains over or underrepresented in the pre-SA corpus compared to the post-SA corpus?

7.5 Methodology

Participants

The participants used in this study are the same as in the previous chapters (4-6). To briefly summarize again, 38 Japanese L1 EFL students travelled to the UK for a 15-day SA experience. The group was relatively homogenous with all participants aged 18-19, female and with proficiency levels around advanced beginner (Common European Framework of Reference (CEFR) Upper A1: TOEIC 230 – 280).

Task administration

A detailed report on the administration of the Lex30 productive vocabulary task has been given in chapters 2, 3 and 4 of this thesis. To briefly summarize, identical versions of the same Lex30 task first used by Meara and Fitzpatrick (2000), were administered to the students on four occasions before, during and after SA. The task was first administered 23 days before the students departed their home country and then immediately before departure for the UK. It was again administered on the third occasion 2 days after arrival back in Japan and finally 21 days after that. Lex30 extracted a sample of up to 120 words from each student in response to 30 cue words on each occasion although many participants produced considerably fewer words than this.

Table 7.2

SA corpora and time points

Time point	1	2	3	4	Total
Types	624	640	854	772	1386
Tokens	1715	1966	2681	2542	8904

Table 7.2 shows vocabulary items gathered at each time point during the SA process in terms of types of word and the total number of tokens. The figures represent the total amount of data produced from the group of 38 participants.

The first research question requires us to investigate how SA corpora can compare with an example of an established reference corpus used by Wmatrix. In this case I have chosen the “BNC sampler written informal” corpus. The “BNC sampler written informal” corpus is one of 12 different reference corpora that Wmatrix provides to assist users. Other reference corpora include BNC Sampler Spoken, BNC Sampler Written, BNC Sampler CG Business, BNC CG Educational and BNC CG Leisure. The “BNC sampler written informal” is used as it may perhaps best represent language produced by Japanese EFL participants. The second part of the first research question will ascertain how our SA data can compare with corpora mentioned by Lin (2014; 2017) in his two studies. To complete both parts of the first question data from all four time points will be grouped together to create a single SA corpus (SAC) totaling 8904 words. This corpus can then be compared with

both the BNC sampler and also look at corpora from Lin (2014; 2017) analyzing for word frequency, keywords, POS categories and semantic domain. This will inform us of the some of the characteristics of language produced by SA participants in general and if a reference corpus can be used appropriately for the purposes of comparison.

To answer the second set of research questions which concern language change during SA, data gathered from the second of four time points will be used as a reference corpus. This data (a total of 1966 words) is produced by the participants at the start of their SA experience shortly before arrival in UK , their new L2 environment. This will then be compared with a target corpus comprising of language produced at Time 3 (a total of 2681 words). Time 3 data was produced by participants just after their arrival back to Japan. Table 4.1 shows the individual SA corpora and data sampling timepoints. Again, as with the first research question, changes in language will be analyzed in terms of keywords, POS categories and semantic domain.

7.6 Results

7.6.1 Comparison of SA data with existing corpora (RQ1)

Word frequency key word analysis

Table 7.3 shows the top ten words that differ most greatly between the corpora. These are words with the largest LL values that are key to the SA Corpus (SAC) as compared with a reference corpus (BNC sampler written informal or more simply BNCI) obtained from the Wmatrix site. In the comparison, items occurring both unusually frequently (positive key words: with a plus sign) and unusually infrequently (negative key words: with a minus sign) as compared to the reference corpora are shown. SAC-F and BNCI-F refer to the frequency with which key words appear in each respective corpus while SAC% and BNCI% show the percentage out of the total number of words contained within each corpus. Most noticeably the words *the* and *of* do not appear at all in the SAC but appear numerous times in the BNCI for the simple reason that the SAC is not discursive. The other eight words on the list appear a number of times (gaining a share of more than 1% out of the total in six cases) in the SAC but barely register as a percentage in the BNCI. An important reason that explains the differences between the two corpora is that the SAC is

hugely influenced by the few cue words used to gather responses whereas the BNCI is derived from a much larger body of data.

Table 7.3

Key word lists: SA corpus versus BNC Samples Informal Written corpus

Rank	Item	SAC-F	SAC%	BNCI-F	BNCI%	+/- use	LL
1.	the	0	0.00	53747	7.21	-	1275.88
2.	white	193	2.17	249	0.03	+	1114.02
3.	apple	136	1.53	21	0.00	+	1084.56
4.	hot	122	1.37	39	0.01	+	9.05
5.	soccer	102	1.15	14	0.00	+	820.59
6.	door	108	1.21	44	0.01	+	777.11
7.	chair	92	1.03	12	0.00	+	742.81
8.	of	0	0.00	26369	3.54	-	625.97
9.	orange	85	0.95	35	0.00	+	610.71
10.	banana	69	0.77	6	0.00	+	571.01

POS keyness analysis

Table 7.4 shows how certain POS categories are over or underrepresented when comparing the SA corpus with our chosen BNCI corpus. The greatest difference that can be seen is that 69.13% of the words belonging to the SA corpus are singular common nouns (Code NN1) as opposed to 15.99% of those in the BNC corpus. A similar pattern can be seen with the base form of the lexical verb (Code VVO) denoting words like *eat*, *write* and *sit*, which account for 8.67% of total words as opposed to only 0.98% in the BNC corpus. In contrast, the SA corpus contains no articles (Code AT) while they represent 7.36% of the BNC. The same pattern is shown with other functional grammar categories such as prepositions or conjunctions which are prevalent in the BNC corpus but virtually non-existent in the SA corpus. These differences clearly demonstrate the differences between the corpora: the SAC comprising of single-word responses to a WAT which are heavily weighted towards the use of nouns while the BNC has discursive text where the grammar categories are more evenly balanced.

Table 7.4*POS: Complete SA corpus versus BNC Samples Informal Written corpus*

Rank	Code	SAC	Freq %	BNC Freq	+/-	LL	POS explanation
1.	NN1	6155	69.13	119215	15.99 +	8377.91	Singular common noun
2.	VV0	772	8.67	7302	0.98 +	1936.15	Base form of lexical verb
3.	AT	0	0.00	54906	7.36 -	1303.39	Article (e.g. the, a)
4.	II	38	0.43	53746	7.21 -	985.92	General preposition
5.	NP1	0	0.00	30541	4.10 -	725.00	singular proper noun
6.	JJ	1573	17.67	62601	8.39 +	679.01	General adjective
7.	NN2	79	0.89	48147	6.46 -	673.10	Plural common noun
8.	IO	0	0.00	26413	3.54 -	627.01	of (as preposition)
9.	CC	0	0.00	25447	3.41 -	604.08	coordinating conjunction
10.	VVN	4	0.04	20956	2.81 -	456.47	Past participle lexical verb

Table 7.5*Semantic domains: SA corpus and BNC Samples Informal Written corpus*

Rank	Code	SAC-F.	%.	BNC-F	%	+/-	LL	Domain
1	Z5	41	0.46	252492	33.86	-	5560.36	Grammatical bin
2	F1	909	10.21	2068	0.28	+	4456.89	Food
3	B5	406	4.56	1623	0.22	+	1612.40	Clothes/personal belongings
4	O4.3	355	3.99	2617	0.35	+	1039.87	Colour / colour patterns
5	Z8	3	0.03	39487	5.30	-	901.10	Pronouns
6	H5	264	2.96	1537	0.21	+	879.57	Furniture/household fittings
7	B1	304	3.41	2432	0.33	+	848.28	Anatomy and physiology
8	O4.6+	172	1.93	701	0.09	+	677.46	Temperature: Hot / on fire
9	X3.1+	71	0.80	0	0.00	+	630.44	Tasty
10	M3	208	2.34	1481	0.20	+	621.57	Vehicles / transport on land

Semantic domain keyness analysis

Finally, Table 7.5 shows the way in which various semantic domains are represented in the two corpora. An important domain associated with the BNC corpus is known as Z5 (grammatical bin) and is made up of words connected with functional grammar such as articles, pronouns, prepositions and conjunctions. This accounts for a third of all words (33.86%) contained within the corpus while with the SA corpus the total is only 0.46%. The F1 (food) domain makes up a substantial part

of the SA corpus with 10.21% while B5 (clothes and personal belongings) and O4.3 (colour and colour patterns) tell a similar story with figures of 4.56% and 3.99% respectively. The same domains for the BNC corpus have negligible values in comparison.

7.6.2 Comparison of SA data with Lin’s 2014 and 2017 corpora

To continue to address the first research question we should briefly look at corpora mentioned in the two studies by Lin (2014; 2017) and compare them with our own SA corpus. Lin (2014) conducts POS analysis on BATTICC, a specialist corpus made of Taiwanese and British participant components. Looking at Table 7.6 we can see that the 11.05% of the Taiwanese participants’ discourse and 7.93% of the British discourse falls into the singular common nouns (Code NN1) category. This compares with share of 15.99% of the BNC corpus and a much higher figure of 69.13% for the SA corpus. The results for the base form of lexical verbs (Code VVO) are similar with regards to proportion with 3.93% (Taiwanese) and 2.82% (British) as compared with 0.98% (BNC) and the top figure of 8.67% for the SA corpus. With POS categories it seems that BATTICC shares quite a similar profile with the BNC corpus but both of these are very different to the SA corpus particularly in relation to the proportions of lexical and functional grammar.

Table 7.6

POS Comparisons: Lin (2014) , BNC and SA corpora

POS	Lin (2014) BATTICC		Present study	
	Taiwan %	British %	BNCI %	SAC%
NN1	11.05	7.93	15.99	69.13
VVO	3.93	2.82	0.98	8.67
Codes	NN1 = Sing. com. noun VVO = Base form lexical verb			

Lin (2017) goes on to analyse four separate corpora BATTICC-O, BATTIC-F, CANELC and CANCODE for word frequency, key words and semantic domains.

Table 7.7

Word frequency and semantic domain comparisons: Lin (2017) , BNC and SA corpora

	Lin (2017) BATTICC				Present study	
	Online	Face	CANELC	CANCODE	BNC	SAC
Word Freq.	I (5.64%) and (3.03%) is (3.02%) my (2.82%)	you (3.17%) I (2.93%) and (2.62%) like (2.53%)	the (3.96%) to (2.39%) a (2.26%) of (1.95%)	the (3.31%) I (2.93%) and (2.80%) you (2.73)	the (7.21%) of (3.54%) and (2.98%) to (2.48%)	white (2.17%) apple (1.53%) water (1.37%) hot (1.37%)
Sem. %						
Z8	16.08		9.10		5.30	0.03
P1	1.97		0.23		0.44	1.79
Z99	2.91		5.48		2.46	0.85
F1	1.91		0.50		0.28	10.21
Code	Z8 Pronouns	P1 Education	Z99 Unmatched	F1 Food		

Looking at Table 7.7 it is clear that there are considerable differences between the SA corpus and some of the corpora described in Lin's (2017) study. Although Lin does not provide complete data sets for all the corpora he mentions in his study (for example, in the cases of BATTICC-F and CANCODE) he still provides sufficient information to highlight some of the main areas. The most frequently occurring noun is *school* which is ranked 14th in BATTICC at (0.96%) (Lin 2017:288). This compares with the SA frequency data where six out of the 10 most frequently occurring words are nouns or adjectives including words like: *white*, *apple*, *water*, *hot* and *door*.

Finally, terms of semantic domains, there are also considerable differences as Table 7.7 also shows. The Z8 (*pronouns*) domain barely registers with the SA corpus (only 0.03%) but makes up 16.08% of BATTICC-O and 9.10% of CANELC. The semantic domain connected with F1 (*food*) has a 1.91% share of BATTICC-O words but only 0.50% of CANELC. Words from the same domain (F1) represent 10.21% out of the total for the SA corpus. P1 (*education*) is quite similar to the SA corpus with percentages of 1.97% (P1) and 1.79% (SAC). Z99 (unmatched) is higher, at 2.91%, but an important reason for this is that many Z99 words (such as proper names, acronyms and abbreviations) were excluded from the SA corpus by special protocol which prohibited their inclusion.

Therefore, in answer to our first question, we can see that there is some difficulty in making a comparison between our SA corpus and (1) “BNC sampler written informal” and (2) with the corpora mentioned by Lin (BATTICC- O and F, CANELC, CANCODE) with respect to the frequency of occurrence of individual words, POS categories and semantic domains. This is not unexpected, seeing as the SA corpus consists of single word responses (rather than discursive text), and is heavily influenced by the semantic domains of the cue words. Given this finding we must now forego our search for a suitable reference corpus and instead attempt to make a comparison between the different components within our original SA target corpus. Our results for the second of our research questions can be seen in the next section.

Table 7.8

Word frequency keyness: SA (Time 2) and SA (Time 3)

Rank	Word	Time 2 Freq.	%.	Time 3 Freq.	%	+/- use	LL
1.	action	5	0.25	0	0.00	-	8.60
2.	lemon	1	0.05	10	0.37	+	6.02
3.	study	4	0.20	18	0.67	+	5.82
4.	pear	0	0.00	5	0.19	+	5.50
5.	jeans	0	0.00	5	0.19	+	5.50
6.	face	0	0.00	5	0.19	+	5.50
7.	break	0	0.00	5	0.19	+	5.50
8.	biology	0	0.00	5	0.19	+	5.50
9.	wonderful	3	0.15	0	0.00	-	5.16
10.	test	3	0.15	0	0.00	-	5.16
11.	information	3	0.15	0	0.00	-	5.16
12.	die	3	0.15	0	0.00	-	5.16
13.	present	0	0.00	4	0.15	+	4.40
14.	past	0	0.00	4	0.15	+	4.40
15.	documentary	0	0.00	4	0.15	+	4.40
16.	culture	0	0.00	4	0.15	+	4.40
17.	wash	12	0.61	6	0.22	-	4.33
18.	fire	5	0.25	1	0.04	-	4.30
19.	fat	5	0.25	1	0.04	-	4.30
20.	foot	3	0.15	13	0.48	+	4.02
21.	bag	9	0.46	4	0.15	-	3.84
22.	people	2	0.10	10	0.37	+	3.63

Table 7.8*Word frequency keyness: SA (Time 2) and SA (Time 3) – continued*

23.	winter	2	0.10	0	0.00	-	3.44
24.	volleyball	2	0.10	0	0.00	-	3.44
25.	three	2	0.10	0	0.00	-	3.44
26.	space	2	0.10	0	0.00	-	3.44
27.	snow	2	0.10	0	0.00	-	3.44
28.	smile	2	0.10	0	0.00	-	3.44
29.	smart	2	0.10	0	0.00	-	3.44
30.	problem	2	0.10	0	0.00	-	3.44

7.6.3 Investigating changes within SA data (RQ2)

Our second research question is more directly concerned with language change during the SA process. Data gathered from SA participants at the second of four time points will act as our longitudinal reference corpus and is compared with data from Time point 3 which will act as our target corpus. Thus we will be able to make a direct comparison with immediately before and after the SA experience. Again, as with the first research question, comparisons will be made in terms of keywords, POS categories and semantic domain.

Word frequency Keyword analysis

Table 7.8 shows the top thirty words (with the largest LL values) that are key to the Time 3 corpus as compared with the reference corpora (Time 2). In the comparison, items occurring both unusually frequently (positive keywords: with a plus sign) and unusually infrequently (negative keywords: with a minus sign) compared to the reference corpora (Time 2) are identified. The table shows individual items which appear more frequently at Time 3 than in the reference corpus (Time 2). More highly ranked words are of particular interest and could be the result of being used more often in an SA environment. For example the word *study* (ranked 3rd) is notable as it made up 0.67% of total words produced at Time 3 as supposed only 0.20% at time two (see Table 7.7). A possibility is that SA participants were required to attend daily EFL classes during their experience and were also perhaps frequently encouraged to *take a break* (ranked 7th) at the same time. As part of their intensive EFL course they learned about British *culture* (ranked 16th) and sometimes

watched a *documentary* (ranked 15th). During their daily life in the UK, *jeans* (ranked 5th) were universally worn but the word *tests* (ranked 10th) occurred at Time 2 but not at Time 3 perhaps indicating that *tests* are not as relevant during the UK SA experience as during the participants' usual life in Japan. Although, given the small sizes of the corpora used in our study, such suppositions may not be entirely accurate they still guide us in considering the influence that an L2 environment may be having on SA learners' language acquisition and knowledge.

Table 7.9

POS frequency keyness: SA (Time 2) and SA (Time 3)

R.	Code.	Time 2 Freq.	%	Time 3 Freq.	%	+/- use	LL	POS explanation
1.	MC	2	0.10	0	0.00	-	3.44	cardinal / neutral number (two, three..)
2.	VVN	0	0.00	2	0.07	+	2.20	p. partic lexical verb (given, worked)
3.	VVG	0	0.00	2	0.07	+	2.20	-ing participle lexical verb (eg.giving)
4.	AT1	0	0.00	2	0.07	+	2.20	singular article (e.g. a, an, every)
5.	RT	1	0.05	5	0.19	+	1.81	quasi-nominal adverb time (now)
6.	VV0	164	8.34	255	9.51	+	1.73	base form lexical verb (e.g. give, work)
7.	NNU1	1	0.05	0	0.00	-	1.72	singular unit of measurement
8.	NN	9	0.46	20	0.75	+	1.56	common noun neutral number (sheep)
9.	DA2	5	0.25	3	0.11	-	1.32	plural after-determiner (few, several)
10.	ND1	1	0.05	4	0.15	+	1.12	sing noun of direction (north, south)
11.	VVI	0	0.00	1	0.04	+	1.10	infinitive (e.g. to give... It will work...)
12.	VM	0	0.00	1	0.04	+	1.10	modal auxiliary (can, will, would, etc.)
13.	VH0	0	0.00	1	0.04	+	1.10	have, base form (finite)
14.	RRQ	0	0.00	1	0.04	+	1.10	wh- gen adverb (where, when, how)
15.	RP	0	0.00	1	0.04	+	1.10	prep. adverb, particle (e.g about, in)
16.	PNX1	0	0.00	1	0.04	+	1.10	reflexive indefinite pronoun (oneself)
17.	MC1	0	0.00	1	0.04	+	1.10	singular cardinal number (one)
18.	DDQ	0	0.00	1	0.04	+	1.10	wh-determiner (which, what)
19.	DA	0	0.00	1	0.04	+	1.10	determiner capable pronom function)
20.	RL	3	0.15	8	0.30	+	1.07	locative adv (e.g. alongside, forward)
21.	VVD	9	0.46	8	0.30	-	0.78	past tense of lex verb (gave, worked)
22.	NN1	1383	70.35	1834	68.41	-	0.61	singular common noun (e.g. book, girl)
23.	RR	12	0.61	12	0.45	-	0.58	general adverb
24.	NN2	17	0.86	28	1.04	+	0.38	plural common noun (e.g. books, girls)
25.	NNT1	12	0.61	14	0.52	-	0.16	temporal noun, singular (day, week)

POS Frequency keyness analysis

Table 7.9 shows the top 25 POS categories (with the largest LL values) that are key to the Time 3 corpus as compared with the reference corpora (Time 2). In the comparison, POS categories occurring both unusually frequently (positive categories: with a plus sign) and unusually infrequently (negative categories: with a minus sign) compared to the Time 2 corpus are identified. The table seems to show that there is only a small difference in the percentage of words belonging to particular POS categories between Times 2 and 3. By far the largest POS category, the singular common noun (Code NN1), shows little change. This slightly decreases from 70.35% to 68.41%. Other category examples such as *base form lexical verb* (Code VVO) and *past tense of lexical verb* (Code VVD) show similar patterns of change.

Semantic domain frequency keyness analysis

Table 7.10 shows the top 30 semantic domains (with the largest LL values) that are key to the Time 3 corpus as compared with the reference corpora (Time 2). In the comparison, semantic domains occurring both unusually frequently (positive categories: with a plus sign) and unusually infrequently (negative categories: with a minus sign) compared to the time two corpus are identified. The table shows that certain domains experience an increase between Times 2 and 3. Some notable examples are T1.1.2 (*time: present; simultaneous*) comprising of words like *everyday, now* and *present* and P1 (*education*) where more than double the number of words were produced at Time 3 than at Time 2 including *school, study* and *test*. The SA participants were staying with host families during their experience and attending daily classes at language school which may have had some influence on the responses produced. It was likely that there was greater than usual need to check appointment times, make requests and use classroom language. An increase in the X7+ (*wanted*) domain with words like *want* or *wish* might perhaps reflect this. Appendix 14 shows the complete results of the semantic domain frequency keyness analysis.

Table 7.10*Semantic frequency keyness: SA (Time 2) and SA (Time 3)*

R. Code	Time 2 Freq	%	Time 3 Freq	%	+/- use	LL	Semantic Group
1. L1-	4	0.20	0	0.00	-	6.88	Dead
2. T1.1.2	2	0.10	14	0.52	+	6.79	Time: Present; simultaneous
3. P1	23	1.17	58	2.16	+	6.72	Education in general
4. X7+	6	0.31	24	0.90	+	6.70	Wanted
5. S2	7	0.36	24	0.90	+	5.33	People
6. B2	3	0.15	0	0.00	-	5.16	Health and disease
7. A8	0	0.00	4	0.15	+	4.40	Seem
8. Z99	11	0.56	30	1.12	+	4.24	Unmatched
9. A2.1+	2	0.10	10	0.37	+	3.63	Change
10. X2.3+	2	0.10	0	0.00	-	3.44	Learning
11. S7.4+	2	0.10	0	0.00	-	3.44	Allowed
12. N3.6	2	0.10	0	0.00	-	3.44	Measurement: Area
13. I1.3+	2	0.10	0	0.00	-	3.44	Expensive
14. X2.4	0	0.00	3	0.11	+	3.30	Investigate, examine, test
15. N5	0	0.00	3	0.11	+	3.30	Quantities
16. I1.1	0	0.00	3	0.11	+	3.30	Money and pay
17. M7	10	0.51	26	0.97	+	3.27	Places
18. L3	59	3.00	58	2.16	-	3.12	Plants
19. X4.2	4	0.20	1	0.04	-	2.98	Mental object: Means, method
20.T1.1.1	1	0.05	6	0.22	+	2.58	Time: Past
21.A5.1+	11	0.56	7	0.26	-	2.57	Evaluation: Good
22.L2	32	1.63	29	1.08	-	2.54	Living creatures
23.A13.3	5	0.25	2	0.07	-	2.43	Degree: Boosters
24. S9	2	0.10	8	0.30	+	2.23	Religion and the supernatural
25. A5.1-	2	0.10	8	0.30	+	2.23	Evaluation: Bad
26.X5.2-	0	0.00	2	0.07	+	2.20	Uninterested/bored/
27.W2-	0	0.00	2	0.07	+	2.20	Darkness
28. S3.1	0	0.00	2	0.07	+	2.20	Personal relationship: General
29. S1.2.4-	0	0.00	2	0.07	+	2.20	Impolite
30. N5.2+	0	0.00	2	0.07	+	2.20	Exceed; waste

7.7 Discussion

As predicted the results clearly show that our SA corpus, which is based on WAT data, bears no comparison with corpora which are created from naturally

produced discourse. Differences in the frequency of individual words, POS categories and semantic domains leave little room for doubt. Particularly in the case of individual words it would be very likely that their corresponding semantic domains will be strongly affected or skewed by the cue words used in the Lex30 task. Perhaps the most noticeable difference can be seen with POS categories where nearly 70% of the words belonging to the SA corpus are singular common nouns as compared with just 16% in the BNC sampler written informal corpus and even fewer with BATTICC (Lin, 2014). At the same time, the SA corpus contains few words belonging to functional grammar POS categories such as prepositions or conjunctions. Semantic domain frequency differences further support these findings. The most frequently occurring domain within the BNC is Z5 (*grammatical bin*) which contains nearly 34% of all words associated with the corpus. Z5 is made up of words connected with functional grammar such as articles, pronouns, prepositions and conjunctions. With the SA corpus the same Z5 category only accounts for less than 0.5%.

With the corpora that Lin describes (Lin, 2014; 2017) the differences are also clear. POS analysis on BATTICC shows that the singular common nouns category only comprises between around 8 to 11 % accounting for an even smaller share than the BNCI. With other POS categories it seems that BATTICC is more comparable with the BNCI corpus but both of these are very different to the SA corpus particularly in relation to the proportions of lexical and functional grammar within each. Only in the comparison of a few semantic domains concerning food, places and education can we detect some similarities. For example, with P1 (*education*) the BATTICC and SA corpus figures are 1.97% and 1.79% respectively. In such cases it could be conjectured that different SA experiences might be influencing participants in similar ways.

The same results reveal an important issue relating to our SA corpus. Words which make up the corpus come from a Word Association Task (WAT), which is basically a wordlist of individual items which can be heavily dependent on the cue words used. The original purpose in using a WAT was to provide a representative sample of vocabulary so that some assessment could be made of an individual's productive vocabulary knowledge. The evidence that we have examined earlier in this thesis (in chapter 4) suggests that this is indeed the case.

Table 7.11*Selection of Lex30 cues and frequently produced associate words*

Cue	Frequently produced associate words (% of SA corpus)			
attack	volleyball 0.48%.			
close	door 1.21%			
cloth	skirt 0.51%	shirt 0.47%	pants 0.46%	
fruit	apple 1.53%	orange 0.95%	banana 0.77%	strawberry 0.63%
furniture	chair 1.03%	house 0.47%	desk 0.43%	table. 0.42%
kick	soccer 1.15%.	box(ing) 0.55%		
pot	water 1.37%	tea 0.27%		
potato	fry (fried) 0.55%	vegetable 0.54%	chip(s) 0.46%	
rice	white 2.17 %	curry 0.21%		

However, Table 7.11 shows that some of the cue words used in the original Lex30 WAT can lead to certain semantic domains being over represented. Examples of such cue words can be seen in the left hand column. In particular cue words like fruit, potato and rice are likely to explain the predominance of the F1 (food) semantic domain (10.21%) within the SA corpus. Other categories include B5 (*clothes and personal belongings 4.56%*) with words like *skirt, shirt* or *pants*, O4.3 (*colour and colour patterns 3.99%*) with words like *white, black, yellow* and *green*, B1 (*anatomy and physiology 3.41%*.) with words like *hand, sleep, foot* or *mouth* and finally H5 (*furniture and household fittings 2.96%*) with words like *chair, desk table* or *bed*. Some cue words used in Lex30, like *fruit* and *furniture* for instance, are hypernyms and are more likely to invite responses from the same domain. The issue concerning semantic domain is further complicated by the fact that some more common words produced by SA participants, fall into more than one category. As mentioned in our explanation of the workings of Wmatrix (section 7.1) the contents of categories are not always mutually exclusive and it is possible for words to fall into more than one of them. So there are cases where words like *apple* or *orange* which belong the F1 (*food*) category can also, at same time, be described as L3 (*plants*). From our Wmatrix analysis it is clear that there are structural differences between the SA corpus used in this chapter and typical reference corpora (including the BNC-derived one used in this case). For the purposes of this thesis using two separate corpora –

one using data gathered from the same participants before SA and the other using data gathered after – proved more informative.

Looking at the second of our research questions we can see comparing corpora produced before and after an SA experience using Wmatrix can inform us of several things. Firstly, it tells us that only slight changes take place regarding the POS categories. Rather than taking this as a sign of SA participants' lack of progress of in mastering grammar it is more likely that it is because of the way the SA corpus was created: by using a WAT. Secondly, semantic frequency keyness analysis seems to tentatively reveal influences a L2 environment may have on SA participants' knowledge and acquisition of language. Referring to Table 7.9, there are a number of domains, possibly associated with SA experiences, which appear to have increased in representation. Examples of some of these include T1.1.2 (*Time: Present; simultaneous*) and P1 (*education*) and where participants use words like *everyday*, *now*, *study* or *test* to discuss daily news and habits with their and host families and teachers. There is also an increase in Z99 (*unmatched*). This is interesting because this domain represents more unusual words that are not in Wmatrix's database dictionary such as *restroom*, *yummy*, *stinky*, *chopstick* and *noodles*. *Restroom* is a word used in the American English normally taught in Japanese schools meaning toilet, *yummy* and *stinky* are colloquialisms perhaps picked up from host family conversation and the last the last two, *chopstick* and *noodles*, tend to be used more within a Japanese cultural setting.

Finally, Lin (2017) explained that certain aspects of his participants' Taiwanese and British cultural background are also evident when comparing key items within semantic domains. This also seems to hold true for Japanese SA participants. Within domains such as F1 (*food*), they tended to produce words more associated with Japanese cuisine such as *rice*, *tofu* or *noodles* and with K5.1 (*sports*) the frequently occurring words featured sports common in Japan like *baseball*, *volleyball* or *karate*. Supporting Lin's findings (2017), the choice of words within each semantic domain seems to demonstrate cultural and social differences with regard to the lexical choices of Japanese participants.

7.8 Conclusion

This chapter has shown that Wmatrix can be a useful tool in comparing corpora produced by SA participants. It has allowed us to test our prediction that the

application of this tool on SA data can help identify certain specific characteristics in the kind of language SA participants are capable of producing and how these might change over time. In particular, results seem to suggest that there are important changes in some of the semantic domains that individual words belong to and this might well be due to participants' response to the changing SA environment. At the same time there is evidence to suggest that, within certain semantic domains, they are also likely to choose individual words which are indicative of their own cultural background.

This chapter has revealed some challenges. In previous studies we saw that it is possible to use reference corpora to make our comparisons (Bianchi, 2017; Willis, 2017; Lin, 2017) but this is not always the case. With a corpus that is more unusual, whether connected with suicide notes (Shapiro, 2011), political manifestos (Willis, 2017) or SA participants' vocabulary knowledge gathered using a WAT, using an appropriate reference corpus may not always yield helpful results. Conducting an analysis on an SA corpus compiled from WAT data only seems possible if we subdivide it into two or more separate corpora so that like-for-like comparisons can be made. However, with the small scale study conducted in this chapter, the limited size of such corpora must also be considered. Comprising of responses produced only at Time 2 and Time 3 by only a limited number of participants there is still a need to be cautious with any conclusions that are made. In spite of this drawback what the study does achieve might be to increase our awareness of possible new research directions.

In future studies there are a number of ways in which it may be possible to improve our analyses of WAT-based corpora. One of these may involve creating our own reference data source by administering an identical WAT to other categories of learner or L1 English speakers which may then be compared with the results from our SA participants. Another might be to change the basic design of our WAT so that it incorporates even more cue words which would then be able to cover a much larger range of semantic domains. If such changes are made it could help further improve our knowledge on how linguistic and cultural influences of a particular L2 environment can contribute towards a better understanding of the design of future SA programmes.

Chapter 8: Discussion

This thesis examines changes in Study Abroad (SA) participants' vocabulary knowledge during their SA experience in terms of word frequency and the characteristics of words they produce. After collecting data from SA participants at four time points using the Lex30 task, four sets of analyses were conducted to investigate changes in frequency, collocation, orthography and semantic domain of words produced. In carrying out such analyses I have attempted to show how different aspects of lexical knowledge might gradually develop over time and be accelerated by the SA experience. This discussion will consider the main findings of these experiments and suggest some directions that future research may take.

The discussion that follows is divided into five sections. The first section (8.1) will reflect on the Lex30 task which was used to collect a substantial sample of vocabulary (some 9000 words from 38 participants) at four timepoints during the main longitudinal experiment. It will consider the changes in both the total number of words and the number of infrequent words produced during the SA process. It will also scrutinize the question of whether the use of narrower word frequency bands can reveal additional changes.

In the second section (8.2), I will consider how the vocabulary samples collected were used to measure changes in collocational knowledge. Although Lex30 was never intended to test such knowledge, the data gathered using such a tool still seemed sufficient to be able to carry out an analysis. I will look at results that I obtained to consider if SA participants indeed acquire a greater degree of collocational knowledge due to their SA experiences and whether this might depend on their level of proficiency (see section 5.5). Comparisons will finally be made with Alqarni's (2017) findings where specifically designed collocational knowledge tests were used on Arabic L1 EFL learners experiencing SA in similar circumstances.

The third section (8.3) will discuss how SA programmes tend to focus more on the improvement of speaking skills rather on the development of writing skills, in particular, orthography. As well as evaluating if spelling accuracy does actually improve during SA (see section 6.4) I will also reflect on whether there is any link between spelling accuracy and the development of speaking (especially pronunciation) skills, which might go some way to supporting findings by Okada (1999; 2002; 2004). Evaluating differences in spelling ability by tracing changes in

identical words produced at multiple time points may bring similar results to those found by Fitzpatrick (2012). Finally, the tendency of different L1 groups to demonstrate certain spelling characteristics will be further considered.

The fourth section (8.4) will reflect on the changes in the kinds of words produced, and the Part of Speech (POS) to which they belong, during both SA and normal EFL studies (see section 7.5). In particular it will consider the role Wmatrix might play in developing more effective teaching materials in preparation for future SA programmes. Are specific words and their corresponding semantic groups any different when they are produced by SA participants compared to regular learners during regular classes in their home country? Considering the different answers to this question might help us decide if learners are adequately prepared in terms of vocabulary knowledge when they travel abroad.

In the fifth section (8.5) I will assess the effectiveness of the measurement tool, Lex30, which has been used to collect data for the four analyses carried out in this thesis. Lex30 has been shown to be a useful and valid test of productive vocabulary knowledge being easy to administer and score and taking less time than equivalent tests measuring the same construct. It allows test takers to demonstrate the breadth of their vocabulary knowledge without constraint in a short time and also distinguish between lower and higher levels of proficiency. Along with these advantages, however, there is still some room for the improvement of this measurement tool. I will suggest some ways in which it can be made even more effective including reconsidering some of the original cue words used and proposing different ways to analyze their responses.

At each stage of the discussion, I shall review the answers to the original research questions and consider the implications for both practitioners (organising SA) and researchers. What the results might mean for the main areas of vocabulary development during SA will be considered and areas of vocabulary development most sensitive to testing will be identified. A significant motivation for my work was to demonstrate quantitatively that SA can promote language development and in this discussion section one of the main aims to see if this has been achieved. The findings reported in this thesis were re-framed and revised from original plans in the light of the COVID-19 pandemic. The pandemic caused the cancellation of SA programme for at least a two-year period and its impact on the future of SA is still to be fully realised. Some consideration will be made of the short and possible long-term effects

of pandemic on SA, and how this might affect language development and particular vocabulary acquisition as part of that SA experience.

8.1 Productive vocabulary knowledge and Lex30

Use of the Lex30 task has shown that it is easy to administer and score, taking less time than either the PVLТ or the LFP (see sections 2.7.1 and 2.8.1 in chapter 2), allows test takers to demonstrate the breadth of their vocabulary knowledge without constraint in a short time and allows for comparison among learners. The replication study reported in chapter 3 lent further weight to support for the validity of Lex30 (see also Fitzpatrick & Clenton 2010; Walters, 2010; etc), indicating a significant change in vocabulary produced before and after an intensive study period. It seems as if Lex30 has indeed become “a robust enough measuring tool to fill an important gap in the battery of tests currently available” (Fitzpatrick & Meara, 2004, p.72; also see section 2.8.3).

8.1.1 Changes in the number of words produced

Regarding the longitudinal experiment in chapter 4, Lex30 also appears to provide convincing answers to two out of the three research questions originally posed. The experiment investigated three research questions including (1) if there was any change in both the total number of words and (2) the number of infrequent words that SA participants produce before, during and after an SA experience. The third (3) research question relating to fine-grained analysis of word frequency will be discussed in the next section (8.1.2). The results showed that there was a significant increase in both the total number of words and the number of infrequent words produced between Time point 2 and Time point 3. In both cases a p value of $p < .001$ was obtained. Regarding the other time points the results were mixed. Between Time points 1 and 2 there was a considerable but not significant increase for the total number of words produced ($p = .015$) while the increase was lower for the number of infrequent words which was again not significant ($p = .353$). Between Time points 3 and 4 there was decrease in the total number ($p = .674$) and in the number of infrequent ($p = .179$) words in both cases. The graphs Figure 4.3, p.90 (for the total number of words) and Figure 4.4, p.91 (number of infrequent words) show these changes. Further graphs (Figure 4.1 and Figure 4.2, p.86) show the results in each case for individual participants. Some reasons for the significant increase between

Time points 2 and 3 have been previously mentioned (see section 4.5.1). It seems likely that immersion in an L2 environment provided greater opportunities for language learning. The result, that is the significant increase in infrequent words between Time points 2 and 3, is similar to those obtained in the earlier study by Fitzpatrick and Clenton (2010) and the results of the replication study both of which are described in chapter 3. The former study had almost the same number of participants (N=40 compared to N=38) and a significant difference between pre and post-tests ($p < .001$). However, the SA duration was considerably longer at six weeks as opposed to 15 days. The replication study had the same number of participants (N=38) and a significant difference in test scores ($p < .001$). In this case the duration was exactly the same.

What could be the reason for the significant increase in the number of infrequent words produced in all three experiments? It is probably safe to assume that short-term SA does not allow time for a large number of words, encountered for the first time, to be learned by participants. However, short-term SA does allow sufficient opportunity for many words that were once part of a learner's purely receptive knowledge to be reactivated and become part of their productive knowledge. Meara (2005) describes an example where spontaneous vocabulary reactivation has taken place as a result of a short period of immersion in an L2 environment (see section 2.6.1). Meara concludes that an L2 environment can play an important part in activating and maintaining vocabulary (something which he refers to as the 'Boulogne Ferry effect'). He points to evidence that active vocabulary can rapidly increase in size upon arrival in an L2 environment and it appears that the same process may well be happening with our Japanese L1 SA participants.

Another important change can be seen in Figure 4.4, p.91, where there is a non-significant rise in productive vocabulary knowledge between Time 1 and Time 2. This concerns the period starting 23 days before the SA programme departure until the departure date itself. As previously suggested this rise was likely due to SA participants undertaking preparation activities such as conversation skills classes or private home study (see section 4.5.1). It could also be due to a practice effect of the test but the stability of scores noted by Fitzpatrick and Clenton (2010) in their parallel forms and test-retest experiments (pp. 540-542) indicates any practice effect was minimal. It might be interesting to further examine this area to see what kind of preparation activity has the greatest impact. One approach perhaps might be to

conduct a questionnaire survey just before the departure date to ascertain whether certain types or length of preparation might have some impact on the number of words that SA participants are able to produce. If any particular form of preparation is found to be more effective then this would have implications for the design of future SA programmes. With the 38 participants in our experiment, the preparation process for their SA experience lasted some six weeks. In this period weekly 90-minute meetings were held during which there were two English conversation classes dealing mainly with the language of requests, London and UK orientation sessions (held in Japanese) and a single English writing class where participants wrote letters to their host families. It might be interesting to contrast this with other forms of preparation used in other SA programmes and whether differences between them these might have a similar impact on Lex30 pre-departure scores.

A further point can be made about the changes in the number of words produced during SA. A review of research has so far found no instance where a delayed post-test has been used with Lex30. In the experiment in chapter 4 we can see the gradual (although not significant) decrease in the number of words produced between Times 3 and 4 and this might be accounted for by the process of language attrition. The delayed post-test (Time 4) seemed to reveal a decline or attrition in vocabulary knowledge after the SA experience has taken place. Research by Ecke and Hall (2012) which has studied this process using other measurement tools, has previously been mentioned (see section 4.1). Ecke and Hall looked at learner's vocabulary knowledge attrition by conducting a case study into rates of vocabulary attrition of a multilingual speaker, showing that even a speaker's L1 can undergo mild attrition when competing with more recently learned languages. Snow et al. (1988) and Schmitt (2010) distinguished between the attrition of receptive and productive mastery of lexical terms and presented evidence to support the view that receptive knowledge does not attrite so dramatically.

The rate of attrition of vocabulary might also be possibly connected with proficiency level which suggests that learners with larger vocabularies might retain their knowledge more effectively over time. Hansen et al. (2002) found the larger the lexical network retained, the greater the chances of reactivating successful links to old words and the greater the chances of having the relevant infrastructure in which to integrate new words. The relatively low proficiency level of the 38 participants involved in our experiment might perhaps then make detection of such a process of

language attrition seem more likely. At the same time there is evidence to suggest that it even might be possible to reduce the rate at which vocabulary is forgotten. Another study by Weltens et al. (1989), looked at the long-term retention of French by Dutch students and actually observed that only very low rates of attrition took place. They explained this by “general cognitive maturation, further academic training, and continued learning of other foreign languages” (Weltens et al., 1989, p. 214).

It might be useful to know if process of attrition continues at the same rate over time and whether different forms of language knowledge are affected more than others. This may help us to decide whether short-term SA programmes can actually have a long-term benefit for participants. Weltens et al. (1989) looked at Dutch high school students learning French. Several years after the students had completed their studies only some lexical and grammatical knowledge showed any signs of attrition while listening and reading skills were still retained. Research by Bahrick (1984) reveals further interesting findings. He examined students who had studied Spanish at high school or university between one and 50 years previously. His analysis showed that while knowledge declined exponentially for the first three to six years, the level of retention thereafter remained unchanged. It appears that the process of language attrition affects some areas of language knowledge more than others and that the process does not tend to continue indefinitely. In evaluating if there is any process of attrition taking place in SA language knowledge, future research might want to consider:

1. conducting further delayed post-tests beyond Time point 4 to see if there is a continuous downward trend in knowledge or if there is a certain point after which it is retained. A comparison of each delayed post-test performance with performance at Test time 2 (before SA) would ascertain the legacy of the benefit of SA (particularly if the effect of any input after the SA period could be controlled for).
2. if further studies can be conducted to see ways in which any attrition in knowledge can somehow be arrested through further language study. Or, alternatively or as Weltens et al. (1989) suggest, a series of remedial activities could be designed to preserve and maintain gains in knowledge attained during SA.

3. if further studies can be conducted to see if particular categories of vocabulary that are lost more quickly (following Snow et al. (1988), who suggest that productive forms and grammar are forgotten more quickly than other skills after a Spanish immersion programme).

8.1.2 Fine-grained analysis of word frequency (RQ3)

The third (3) research question investigated if a finer-grained application of word frequency analysis using narrower frequency bands could make any contribution to SA vocabulary acquisition research. Looking carefully at our results and Figure 4.5 (p.93), adopting a fine-grained approach does not appear to reveal much further information. What we might have expected is a tendency for SA participants to produce gradually fewer words within each successively less frequent word band. However, the frequency profile for the first 500 most frequently occurring word band looks similar to the next (501 to 1000 frequency word band). This suggests that within the first 1000-word group (>1K) there was little change in the frequency of word use. There does, however, seem to be an increase in the number of words in the 501 to 1000-word frequency band at Time 2 but this is very slight. What is more evident is the dramatic decrease in the number of words used within the 1001 to 1500-word frequency band which we might expect with lower proficiency learners. This is followed by a further clear decrease within the 1501 to 2000 band. From this point it becomes difficult to detect a regular pattern. The number of words produced in the 2501 to 3000-word band seems to slightly exceed the number produced within the preceding 2001 to 2500-word band which seems to lend support to the idea that learners seem to be able to acquire and produce words roughly according to their frequency during earlier stages in the learning process but this ability becomes less predictable the further we proceed along the frequency continuum.

The fact that lower frequency word bands after the 2000 mark do not seem to follow a similar pattern of improvement seems to confirm Aizawa's (2006) and Milton and Alexiou's (2009) findings that lower word frequency bands (particularly after 3-4000 words) lack usefulness (see section 4.5.3). With the data gathered in our experiment it seems that as words become less frequent, our SA participants are less likely to acquire them in a strict predetermined order.

One further point can be made. Some of the later lower frequency word bands contain an unusually high number of words (see Figure 4.5, p.93). This is even allowing for the fact that the frequency word bands proposed by Kremmel (2016) become progressively wider (in 500 words steps from 0 to 3000 words, 1000-word steps between 3001 and 6000 words and in 2000-word steps thereafter). For instance, within the 2001 to 2500 frequency band, SA participants produced a total of 403 words which is considerably more than the 235 words produced within the previous 1501 to 2000-word band. Furthermore the 3001 to 4000-word band also contains more words than one might expect. This seems to confirm Aizawa's (2006) findings that lower word frequency bands lack usefulness as they do not seem to follow a steady pattern of improvement seen with higher frequency ones (see section 4.5.3). However, Aizawa argued that this was the case only after the 3-4000 frequency word bands and not with the higher frequency ones as in this case. A possible explanation is that some of our SA participants are demonstrating that they are capable of producing a small number of infrequent specialist words which pertain to their field of study. This becomes more evident when one takes a closer look at individual words like *chemistry* (ranked frequency 2093rd), *biology* (2319th), that were produced in response to the cue word **science** and *stroke* (3357th) and *protein* (3476th) in response to **substance**. The fact that such less frequent words appear a number of times seems more likely given that 25 out of our 38 SA participants are studying nutrition as a major part of their university course (see section 4.3.1).

8.2 The effects of SA on collocational behaviour

The experiment reported on in chapter 5 seems to provide affirmative answers to the two research questions that were asked. These were (1) whether there is any change in learners' productive knowledge of collocations with an exposure to an L2 environment and (2) whether certain proficiency groups were favoured over others.

8.2.1 Increases in productive collocational knowledge

The investigation found that in many cases learners' production of collocations increased with exposure to an L2 environment. Previously, in chapter 2, I focused on possible changes in formulaic language including collocation which may occur during an SA experience. I mentioned studies by Möhle and Raupach (1987), Towell et al. (1996), Regan (1998) Foster (2009) and Siyanova and Schmitt

(2008) which revealed how SA learners were able to produce more fluent natural sounding language and demonstrate improvements in lexical organization (see section 2.5.2).

Looking again at Figure 5.1 and the accompanying Table 5.1 (see p.114) we can see that the mean number of collocations produced by each participant increased from Time point 1 to Time point 2 (during the three-week period before SA) and again from Time point 2 to Time point 3 (during SA itself). Between Time point 3 and 4 the mean number of collocations declined slightly. A repeated measures ANOVA indicated a significant increase in the number of collocations produced in the period before and during the SA experience and that the slight decrease seen on the participants' return to their home country was not significant. In answer to the second research question of whether any proficiency group was favoured we can again look at Figure 5.1. This shows that the higher proficiency learners (N=19) produced more collocations per individual at each time point than lower proficiency learners (N=19). Although Figure 5.1 shows an increase a repeated measure ANOVA analysis indicates that the increase is not significant either in the period before SA or during the SA experience itself. ANOVA analysis of lower proficiency learners revealed a different result. This indicates a significant increase in the number of collocations produced in the period before and during the SA experience and that the decrease seen on the participants' return to their own country was not significant.

These results support Alqarni's (2017) view that learners, particularly lower proficiency ones, seem to acquire collocational knowledge more easily than those with higher level proficiency. Alqarni found that as the SA participants continue to spend longer in an L2 environment their overall proficiency increases, which is to be expected, but at the same time the rate at which they acquire collocational knowledge slowly declines. The results of the experiment in chapter 5 also support Fitzpatrick's (2012) own findings of position-based (syntagmatic) collocational responses. She discovered that there was "a detectable – though very slight – upward trend in the position-based responses" produced by her case study subject (2012, p. 91). In a similar way to Alqarni (2017), she found that the greatest increase in the rate of acquisition of productive collocational knowledge took place at an earlier stage of SA after which it seems to plateau for the remaining period (Fitzpatrick, 2012, p. 91).

The findings suggest that lower proficiency learners are more favoured, meaning that they are more likely to acquire productive knowledge of collocations at a faster rate than higher proficiency learners. This goes some way to support Gobert (2007) who discovered that higher proficiency Saudi ELF learners studying at home (AH) had a relatively low level of collocational knowledge when compared with lower proficiency learners carrying out SA studies despite having received specialized instruction on collocational use. Gobert concludes that an L2 environment can be important for helping lower proficiency learners in particular, suggesting that collocational knowledge might be more easily acquired by exposure to an L2 environment rather than being learned through formal instruction. Therefore, with the SA participants taking part in the experiment in this thesis, it might be the case that the changes in collocation knowledge are more easily detectable due to their relatively low proficiency level. With a learner group consisting of higher level proficiency participants, undertaking a similar short term programme, any detection of change might prove more difficult.

8.2.2 Creating new ways to measure collocations

I found that using the Lex30 productive vocabulary task was an effective way to gather a sufficiently large sample of vocabulary with which to investigate learners' abilities to use collocations. It offers an alternative to previous methods such as those devised by Bahns and Eldaw (1993) or Bonk (2000) which used measures such as sentence-translation and cloze tasks, to collect data on productive collocation knowledge. My investigation in chapter 5 suggested some increase in the production of collocations although two limitations can be noted. Firstly, defining collocations remains problematic (as it was for Brown (2018); Barfield (2009); etc) because using dictionaries or L1 data does not always capture experience-specific or culture-specific collocations. Some of the English language collocations learners produce have validity within the Japanese context but, because they have Japanese cultural references, are not typically found in collocation dictionaries or among L1 speakers of English (with the possible exception of those based in Japan). Some examples of ways in which this might happen are with cue words like *attack* and *rice* producing culturally influenced responses rather than collocates. It seems clear many words produced by Japanese L1 EFL learners in the experiment are not a demonstration of

L1-like collocational knowledge but simply show how L2 cues can stimulate them to produce a variety of L1 inspired cultural concepts and expressions.

Secondly, the number of collocations produced was limited - they are only one of a number of possible types of word association. A possible way to encourage an increase in the number of collocations SA participants is to consider previous attempts at their measurement and adapt them. Barfield (2009), introduced *LexCombi*, which is similar to Lex30 with 30 carefully chosen words which required three (as opposed to four) associations for each. The main difference with Lex30 is that the measure asked specifically for three collocations whereas previously there was no requirement made for the kind of responses that subjects were expected to produce. Items were scored against a list of what Barfield termed as “appropriate” collocations for each, which was created using the *Oxford Collocations Dictionary* (1st edition) (Crowther et al., 2002) and *Collins Wordbanks Online* (Harper Collins, 2004). Barfield presents LexCombi as a highly efficient way to elicit collocations from learners within a short time. Unlike earlier methods it features neither cloze nor translation tasks and is much more open, allowing learners to produce language in a less constrained manner.

Brown (2018) proposes adapting *LexCombi*'s basic design so that participants do not have to rely on instructions alone to produce a higher number of the desired collocational data. He emphasises that we perhaps place too much faith in instructions, and this might especially important given that many low proficiency learners such as those used in our SA experiment would be unfamiliar with the term “collocation”. Brown's (2018) adapted version of LexCombi directs participants to enter each response either to the left or to the right of the cue. The purpose of this is to encourage participants to think in terms of how words follow one another and guide participants towards providing specific collocational responses. He also reviewed the explanation of collocations given in the original instructions and informed participants that they should use “words that you would use either before or after the cue word”. Brown (2018) compared his adapted version of LexCombi with the original and found a number of differences between the two formats. First, the adapted *LexCombi* format resulted in a significantly higher number of overall responses. Second, there he found that there was a slight but insignificant tendency for the adapted format to elicit more responses which were collocations. Finally, the adapted format appeared to encourage participants to produce more single-word

responses rather than multi-word responses which is an outgoing issue with word association tasks.

Given such evidence that an adapted version of a word association task (LexCombi or Lex30 given that they are similar) might bring better results it would be a relatively easy to implement. Figure 8.1 below gives an idea on how such a task might look. Perhaps together with clear bilingual instructions for the benefit of lower proficiency level subjects, use of a new “collocational-purposed Lex30 task” might result in participants producing an overall higher number of words, a higher percentage of which are likely to be collocations.

Table 8.1

Comparison of the original Lex30 task with an adapted version (after Brown 2018, p.101)

Original version

attack	1	2	3	4
board	1	2	3	4
close	1	2	3	4
cloth	1	2	3	4

Adapted version

_____ attack _____	_____ attack _____	_____ attack _____	_____ attack _____
_____ board _____	_____ board _____	_____ board _____	_____ board _____
_____ close _____	_____ close _____	_____ close _____	_____ close _____
_____ cloth _____	_____ cloth _____	_____ cloth _____	_____ cloth _____

8.3 The effects of SA on orthography

The investigation that I reported in chapter 6 asked three main questions. Firstly (1), it asked whether an SA learner’s spelling tended to improve, decline or remain the same over the course of an SA programme particularly where multiple identical responses are given by the same subject for the same cue words. Secondly (2), it tried to find out if lower L2 proficiency subjects were likely to produce more misspelled words in proportion to the total number of words produced than higher L2

proficiency subjects. This is when making the assumption that misspelled vocabulary items indicate some degree of partial vocabulary knowledge while correctly spelled items indicate a more complete knowledge. Finally, (3) it explored the nature of the pattern of spelling errors that subjects made and asked what this might reveal about the impact of their L1.

8.3.1 The importance of spelling

Findings from the study reported in chapter 6 indicate that a general improvement in spelling proficiency takes place during SA. This discussion will focus on the original research questions and compare the findings with past studies. Many short-term SA programmes concentrate on the development of speaking and listening skills as opposed to reading and writing. However, there is case to be made for the improving English pronunciation of many SA participants having a corresponding beneficial effect on spelling. This section will consider a proposal for investigating a link between speaking skill development and spelling proficiency. Correct spelling is such a crucial part of English language learning that its instruction remains an important goal of teachers and schools. With short-term SA it seems to be an area that we may reasonably expect to experience some neglect. However, it could well be the case that speaking and spelling skills are more closely connected and that the acquisition of both can be mutually facilitative and reciprocal.

8.3.2 What orthographic changes were revealed?

The experiment in chapter 6 found that the average participant's spelling improved over the course of the SA programme. The first of the original research questions (see section 6.4) asked if learners' spelling improved, declined or remained the same during SA. Any assessment of spelling accuracy could only be made where the same response was intended each time. The proportion of misspelled words to total words produced at each point by SA participants, gradually declines from Time 1 to Time 3 before rising slightly again at Time 4. The experiment identifies 64 examples of occasions of spelling improvement where the three or four instances of the same word are provided in response to the same cue, appear to be spelled correctly at Time 3 or Time 4. These results tie in with Okada's (1999) findings which showed that Japanese SA participants appeared to make gradually fewer

spelling mistakes than their contemporaries studying in a home environment (as noted in chapter 6, section 6.8).

The second research question asked whether the proficiency level of the learner has any effect on spelling. We can start by making the assumption that misspelled vocabulary items reveal some degree of partial vocabulary knowledge while correctly spelled items indicate a more complete knowledge. With two SA participant groups divided on the basis of proficiency is it then the case that the number or proportion of partially-known words a learner knows is dependent on their proficiency level? More specifically are lower L2 proficiency subjects likely to produce more misspelled words in proportion to the total number of words produced than higher L2 proficiency subjects? In my experiment the 38 SA participants are divided into two proficiency groups using their raw Lex30 scores as a proxy for proficiency. The results indicate that the lower proficiency L2 group produce a higher percentage of misspelled words compared to the higher proficiency group. The lower proficiency group misspelled 15.7% of words at time 1, 12.3% at time 2, 10.8% at Time 3 and 10.5% at Time 4 with an average of 12.0% overall. This compares with the higher proficiency group who misspelled 9.4% of words at Time 1, 8.5% at Time 2, 5.7% at Time 3 and 6.5% at Time 4 with an average of 7.30% overall. Earlier I mentioned that Zareva (2012) (see section 6.2) also found that her lower L2 proficiency group had a greater partial vocabulary knowledge in proportion to complete knowledge compared to her higher L2 proficiency group. Both studies perhaps show that lower proficiency learners are likely to have a higher proportion of partial vocabulary knowledge in their overall lexicon than higher proficiency ones.

The final research question posed in chapter 6 asks whether the nature of the pattern of spelling errors that subjects make is influenced by their L1. The results showed that there are at least three clear indications of L1 influence. Firstly, with consonants, it seems that in many cases Japanese L1 speakers are uncertain whether to use < l > or < r >. Examples include words like *delicious* or *technology* which are misspelled as *derishious* or *technorogy*. Secondly, there many cases of extra vowel insertions, particularly where there are consonated clusters such as with < dr > or < str >. Examples include *dorama* instead of *drama* or *starawberry* instead of *strawberry*. Finally, it seems that Japanese L1 speakers tend to add vowels at the end of words perhaps due to the influence of import loan words or the way that Japanese kana (syllabic alphabet for loan words) is pronounced. Example are words such as

museum-u or *wash-e*. For further details on the influence of L1 on patterns of spelling please refer to section 6.2.4.

8.3.3 How can an SA experience improve spelling skills?

An important question that we can now consider is whether an SA experience can actually improve spelling skills? As previously mentioned, an important emphasis during short-term study abroad programmes has always seemed to be on the development of speaking skills. A survey of prospective Japanese SA participants showed that 86% of them had expectations of making improvements in speaking and listening skills while only 14% expected to make progress in all four skill areas (Matsumoto, 2012). Given the short duration of many programmes it seems there might be fewer opportunities to develop writing skills and studies which investigate their improvement including spelling skills, tend to examine changes in student performance over much longer periods (e.g., Sasaki, 2011). Certainly with the students who provided the data analysed in this thesis, there was little time to incorporate a significant writing component into the syllabus covering their short 15-day stay in the UK. Most of the emphasis was on the development of speaking and listening skills, enabling participants to communicate successfully with their host families and the class teachers.

Tuladhar and Akatsuka (2017)'s work can shed light on some of the findings from my own orthography study reported on in chapter 6. Their research looks at how pronunciation practice, or in other words the development of speaking skills, can lead to improvements in spelling (see section 6.2.4). In the same way the increased exposure to an L2 environment during SA along with opportunities for speaking practice might bring about similar improvements in spelling for participants. The number of categories that Tuladhar and Akatsuka use to classify "commonly misspelled words" however, far exceeds the limited attempts for error classification carried out in the experiment described earlier in this thesis. What their research can do is to help inform a future study on SA vocabulary which specifically looks at changes in orthographic knowledge. For example, it might be useful to classify any misspellings that learners make more comprehensively. This has been done in previous studies by Cook (1997), Gunion (2012) and Okada (1999). The pattern of misspellings revealed in the responses to Lex30 task can also be supplemented by categorizing errors relating to certain phonological and

grammatical issues in the same way as in Tuladhar and Akatsuka (2017). More challenging would be the task of recording and transcribing SA participants' classroom, out-of-class and host family interactions. However, if a substantial SA spoken interaction database could somehow be created it would be relatively straightforward to compare it with patterns of orthographic change taking place during SA programmes.

Finally, the last of the research questions posed in chapter 6 involves the pattern of spelling errors that subjects make which relate specifically to L1 influences. The effects of travelling to an L2 environment, perhaps for the first time, will likely have a beneficial effect on participants' ability to spell more accurately. With increased exposure to target language input could mean that they make greater improvements in pronunciation which, in turn, will lead to better spelling. As mentioned in section 6.6.3, Japanese SA participants might gradually shed some of their L1 cultural influences by learning how to recognize the correct use of consonants (distinguishing between < l > and < r > and also between < b > and < v >), learning to avoid extra vowel insertions (*drama* instead of *dorama*) and by being careful not to add vowels to the end of some words (*museum* instead of *museumu*).

8.4 The effects of SA on the acquisition of vocabulary within semantic domains

The following section will reflect on the vocabulary that SA participants tend to produce and subsequent semantic grouping analysis. The discussion will attempt to cover several areas. It will start by again considering the original research questions (see section 7.4 in chapter 7). Question (1) looked at a corpus of words produced by short-term SA participants and asked whether certain words, and the Part Of Speech (POS) categories and semantic domains to which the same words belong, appeared more frequently or less frequently than would normally be expected when compared with a reference corpus (BNC sampler). Question (2) attempted to identify if there was any change in individual words, and the POS categories and semantic domains to which they belong, at time points before and after an SA experience to assess whether an L2 environment has any impact on learning. After considering the research questions we will look at an analysis of language produced by SA participants with the aim of developing better teaching materials for SA preparation courses in Japan. Some previous English teaching

material surveys that have taken place in Japan will be considered. In some cases these have involved creating corpora and used various methods, including Wmatrix, to examine them with a view to making improvements in the teaching materials used. Finally, some ideas for a study with the aim of developing better SA preparation teaching materials using Wmatrix, will be introduced.

8.4.1 Comparisons of an SA corpus with other corpora

The results of the experiment in chapter 7 seem to go some way to answering the original research questions that were posed (see section 7.4). The first question asked whether SA participants tend to use words which occur more frequently or infrequently when compared to existing language corpora. The overuse or underuse of certain words and the over representation or under representation of particular POS categories and semantic domains can clearly be identified. However, it was also clear early on in the investigation, that there were some fundamental differences between the SA corpus and an existing corpus derived from the BNC. By their very nature the BNC corpora are wholly derived from free discursive speech or writing, whereas the Lex30 task elicits a much more limited range of vocabulary produced by SA participants. For example, the SA corpus would be likely to have a far greater proportion of nouns, verbs, adverbs and adjectives and far fewer discursive (functional or grammatical) words than the BNC.

With POS the results are as we predicted earlier (see Table 7.4, p.164). The greatest difference that we can see is that 69.13% of the words belonging to the SA corpus are singular common nouns as opposed to 15.99% within the BNC corpus. In contrast, 7.36% of the BNC is comprised of articles while the SA corpus has virtually none. The same pattern is shown with other functional grammar categories such as prepositions or conjunctions which are prevalent in the BNC corpus but virtually non-existent in the SA corpus. What the experiment in chapter 7 shows, as far as POS analysis is concerned, is that making any useful comparison between an SA corpus using data collected by Lex30 and a BNC corpus, would be problematic.

It was notable that some items appeared a disproportionately high number of times in the Lex30 corpus. Two key causes of this are likely to be 1) the influence of specific cue words and 2) cultural influence. The reason that such words are produced more often than usual is likely to be due to either the cue words used which are specifically selected to encourage more infrequently occurring words or involve

cultural reasons. For instance the word *white* often occurs in response to the cue word *potato* which, in turn, is often associated by many Japanese people with *rice*. The word *soccer* might be produced more often than normal due to the recent popularity of the sport in Japan among women, who make up the entirety of the participant group involved in the study.

The clear influence of cue words was evident in the semantic domain analysis, too. Many cue words had some connection with food such as *fruit*, *potato*, and *rice*. It so happened that the most commonly occurring domain is FI (Food) which comprises of 10.21% of all words produced by SA participants in comparison to just 0.28% which occur in the BNC. The next are B5 (Clothes and belongings) (4.56% compared to 0.22%), O4.3 (Colour and colour patterns) 3.99% compared to 0.35% and H5 (Furniture and household fittings) 2.96% compared with 0.21%. More details can be found in Table 7.5, p. 164. Although cue words are likely to have some influence it might also be true that in the case of an SA experience, food is likely to be a popular theme especially with young people trying new dishes and tastes for the first time. Also part of the attraction of SA is the fact there are a number of opportunities for shopping and trying new clothes or fashion.

Given the findings noted above, it can be concluded that using a suitable reference corpus from existing sources, particularly when they vary so much in terms in types of word, POS and semantic groups would not be beneficial. Instead it appears to be more useful to make a comparison between the different components within our original SA target corpus.

8.4.2 Comparisons of semantic domain representation

The second research question addressed in the WMatrix analysis (see section 7.4) asked if any of the language that SA participants produce during their experience changes in terms of the words they produce and the semantic domain to which they belong. When making comparisons between Time 2 (the start of the SA visit) and Time 3 (the end), some interesting differences appear as can be seen on Table 7.8 (p.168). For example the word *study* made up just 0.20% (4 in total) of the total number of words produced at Time 2 compared with 0.67% (or 18 in total) at Time 3 representing more than a fourfold increase. Other words that tended to be produced more were *break* (produced on 5 occasions at Time 3 and none at Time 2), *documentary* and *culture* (both produced 4 times at Time 3 and none at Time 2).

These words perhaps reflect the intensive new educational and cultural environment that is being experienced by the SA participants. With regards to the changes in semantic domain the differences are also detected (see Table 7.10, p.171). For example the P1 (Education) domain (P1 is domain that Wmatrix uses to refer to words connected to education) which contains words like *school, study teacher, student, university, college, exam, class* or *homework* made up 1.17% of all words produced at Time 2. By Time 3 this had nearly doubled to 2.16%. Another domain X7+ (Wanted) consisting of words like *want, wish, purpose* or *plan* was 0.31% of all words produced at Time 2 but this increased to 0.36% at Time 3. Finally, the domain Z99 (Unmatched) made of words like *yummy* or *stinky* which are not normally found in dictionaries doubled from nearly 0.56% at Time 2 to 1.12% at Time 3. This suggests again that SA participants' vocabulary is influenced by the environment in which they find themselves and communication with their teachers and host families.

In both this study and the research by Lin (2014; 2017) the words produced reveal specific themes that participants are mainly concerned about. Words connected with school life, school events and sports like *soccer* and *volleyball* are particularly common. With semantic domain category other notable similarities occur. P1 (Education), F1 (Food), X7+ (Wanted) S3.1 (Personal Relationships) and Z99 (Unmatched) are well represented in both studies.

8.4.3 Future applications of Wmatrix to SA research

The methodology and findings related to semantic domains that I have reported in this thesis suggest that WMatrix might be a useful tool in SA research. The analysis reported in chapter 7 reveals that participants tend to produce words belonging to certain semantic categories and when comparing Times 2 and 3 for example, it seems that an SA experience activates more items from certain domains. Examples include words connected with travel, the host family experience and classroom language. Wmatrix has the ability to gather information about which words might be useful to learn in preparation for an SA programme. One way in which it could be used is in the development of study materials for students who are considering SA in the future. A starting point might involve gathering materials such as English textbooks which the same students regularly use in their high school or university classes and subjecting these to Wmatrix testing to ascertain the proportion of words that are associated with certain semantic categories. This information could

then be compared with the Wmatrix findings obtained from SA participants to assess the degree of overlap. If items produced by SA participants do not appear in the course materials, then this might indicate that the materials that learners regularly use in their studies are not preparing them adequately for SA and would therefore could be improved or modified in some way.

Assessment of English textbook and EFL course materials is a complex task. Some Japanese studies have attempted to measure Japanese EFL students' vocabulary knowledge at different stages from elementary school level (age 7) until high school (normally until ages 17 to 18). In the same way Wmatrix can perhaps help with the development of better English teaching materials and extend their use to creating new materials for SA programmes. With elementary school textbooks, Hoshino (2020) investigated the vocabulary range and characteristics of words. Analyzing different publications she noted that a large number of concrete nouns are used which are related to colours, animals and jobs. There also appeared to be more action than static verbs and many more positive, rather than negative, adverbs and adjectives. Hoshino used *CasualConc* (Imao, 2019) to find the number of tokens and word types used then calculated the number and percentage of words for each part of speech category. Hoshino concluded that her research could prove useful for teachers when designing new elementary English courses. Wongsarnpigoon (2018) looked at junior high schools English textbooks using the computer programme *Range* (Heatley et al., 2003). His analysis showed that the textbooks could not provide enough exposure to new vocabulary items for sufficient learning to ensure examination success. Wongsarnpigoon concluded that the use of software like *Range* can inform teachers of materials' lexical content helping them create future textbooks which can provide enough exposure to useful words. Finally, Sugiura et al. (2020) analyzed the vocabulary in English textbooks used by high school students. They used *Compleat Lexical Tutor* (Cobb, n.d.) to create a complete vocabulary profile for each textbook. Their study found that the communicative activities in textbooks did not use vocabulary that was suitable to deepen students' comprehension.

In the examples above corpora consisting of different learning materials were analyzed using different online software (*CasualConc*, *Range* and *Compleat Lexical Tutor*). This allowed the rapid calculation of the number of word tokens, word types and a breakdown of different grammatical categories according to frequency band.

What such software could *not* accomplish is the categorization of words in each corpus into specific semantic groups. This limitation was only overcome in the first case exception where the elementary school materials examined by Hoshino (2020) had a limited range of vocabulary content thereby allowing some degree of manual semantic categorization.

There are few examples to be found where Wmatrix has been used in the field of linguistics to analyze a corpus in order to carry out semantic categorization. Referring to Rayson's Wmatrix website (<https://ucrel.lancs.ac.uk/wmatrix/>) and his list of examples and applications of the software, the author identified only 5 out of a total of 188 where it had been used in some kind of EFL analysis. There are two examples which are relevant for this discussion. Firstly, Nakano and Koyama (2005) who prepared a large corpus derived of words from 263 abstracts of English language mechanical and electrical engineering journals. They used Wmatrix to analyze the corpus concentrating on part-of-speech and semantic tags, and compared the results with those of the British National Corpus (BNC) written corpus sampler. They found differences in frequencies of certain semantic and part-of-speech categories as well as differences in the use of verbal forms and multi-words. Nakano and Koyama (2005) used the most important findings to develop web-based ESP e-learning materials being for engineering graduate students. The second example is Miura (2020) who compared the use of language in essays written by British and Japanese elementary and junior high school students. His study created corpora for each group which each identified the quantitative difference in verb use as well as part-of-speech and semantic categories. He conducted his analysis with the aim of creating more authentic and effective EFL teaching materials for Japanese junior high school students.

There are several reasons why analysis of English textbook and EFL course materials might be necessary but the last two examples show that Wmatrix can play an important role. Studies of English vocabulary materials used in Japanese schools that we have looked at so far: (1) attempt to draw comparisons between the level of knowledge found in officially approved textbooks and that required for a particular examination, (2) try make some comparison between textbook vocabulary to see if it actually meeting the required level of communicative competence or 'target knowledge' officially demanded by MEXT and finally (3) making comparisons between the English used in textbooks and the English used by native speakers in an

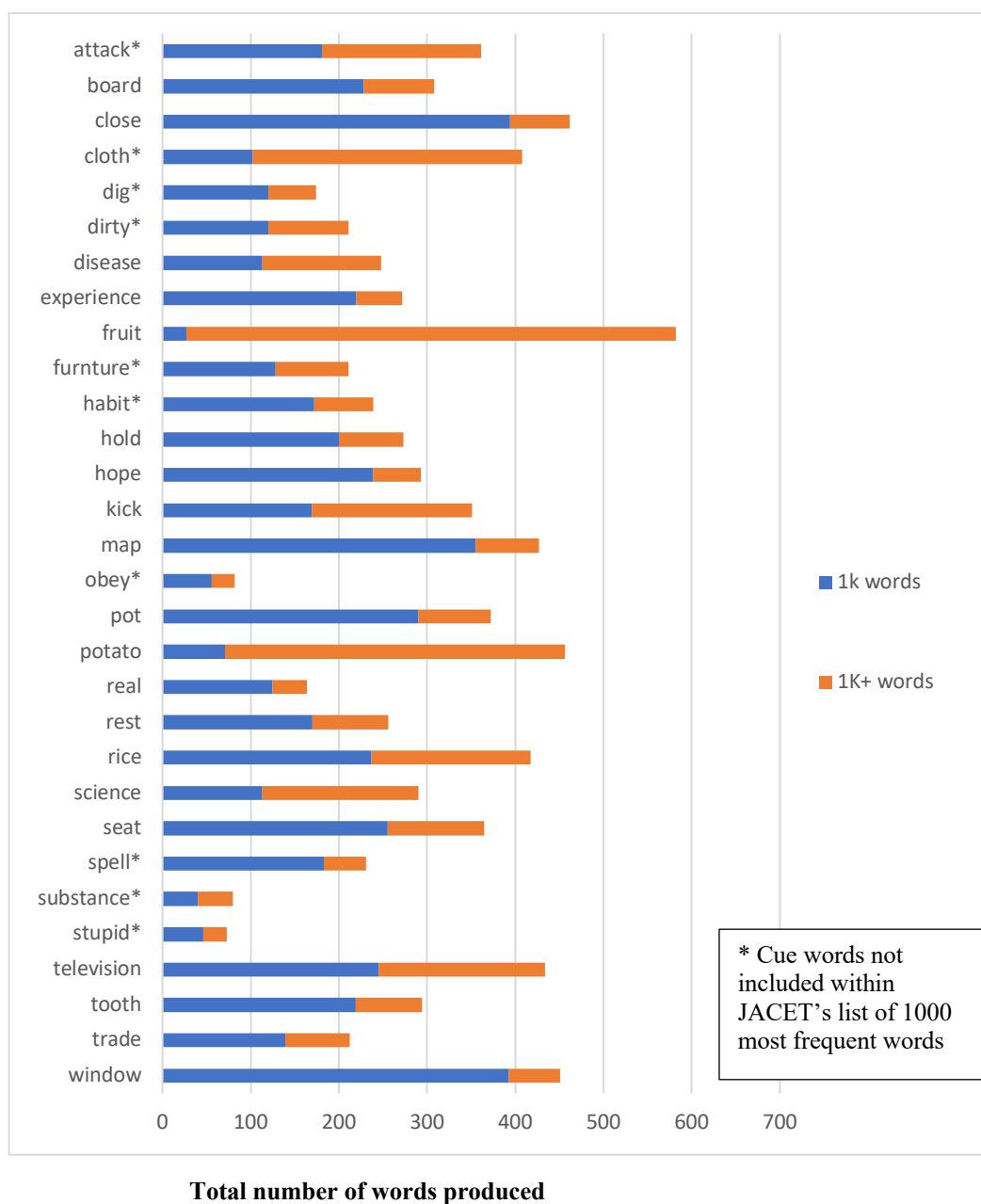
attempt to aid the development of more useful teaching materials. The last point is perhaps the most relevant for developing a project concerning SA. As with Nakano and Koyama (2005) and Miura (2020) Wmatrix might be used in the same way to develop more effective teaching materials, this time to enable more effective preparation for future SA programmes.

8.5 Adapting Lex30 for future research

As a tool for measuring changes in productive vocabulary performance,

Figure 8.1

Number of responses to cue words: 1K and 1K+ frequency levels



Lex30 has demonstrated a number of advantages over existing methods. In the studies reported in this thesis, it has successfully supported the view that an SA period accelerates acquisition of productive vocabulary knowledge. However, a number of limitations to the Lex30 test have become apparent.

It is possible that addressing these limitations might further improve the sensitivity of Lex30 to small changes in vocabulary knowledge over short periods of time. Some important areas might include: (1) accounting for the variation in the number of responses produced by SA participants according to which cue is used, (2) dealing with possible issues with the scoring system and, finally, (3) reviewing the way in which the task is administered. In the following section I present some suggestions that might be made for modifying the Lex30 task. This might have the effect of increasing the number of words available for analysis and perhaps more easily allow us to detect any changes that might occur.

8.5.1 Variation in the number of responses

The first area to consider is that there seems to be a large difference in the number of responses given by SA participants to certain cue words. While some cues attract a large number of associate words, others attract comparatively few. The SA participants in the experiment produced a total of 9449 tokens in response to the Lex30 task over 4 time points (the maximum possible total was 18,240 - 4 responses to every item at every timepoint by every participant). However, there was considerable imbalance depending on the cue word used. Figure 8.1 shows the number of responses, including both frequent (1K) and infrequent (1K+) words, made to the original cue words. What is immediately clear is that the number of words produced in response to particular cues is highly variable. While some attract a comparatively high number of words (e.g., *fruit* 582, *close* 462, *potato* 456 and *window* 451), others (e.g., *dig* 171, *obey* 82, *real* 164, *stupid* 73 and *substance* 80) have far fewer responses. A likely reason for this is that some participants may have limited knowledge of one or more of the cue words and were therefore unable to spontaneously produce responses. An important stipulation made by Meara and Fitxpatrick (2000) in describing the use of an early version of Lex30 is that the cue words should appear within the 1000 most frequently occurring words on Nation's frequency word list (Nation, 1984). They mention that "all the stimulus words are

words which even a fairly low-level learner would be expected to recognize” (2000, p.22).

However, if we look at Figure 8.1 once again we can see that the cues I mentioned previously: *dig*, *obey*, *real*, *substance* and *stupid*, each attracted fewer than 200 responses in total from 38 SA participants. It is of some note that these particular cues were unable to elicit any response from the majority of participants. This might lead one to question whether the use of Nation’s 1984 wordlist might be entirely appropriate with this particular group. Furthermore, the graph shows an asterisk* next to ten out of the thirty cue words used for Lex30. These ten cues appear in Nation’s 1984 wordlist but not within the first 1000 most frequently occurring words on a more recent wordlist, the JACET 8000 word frequency list (Ishikawa et al., 2003). Interestingly, these same ten cue words, upon closer examination, seem to encourage a fewer than average number of responses overall perhaps again underlining the need the review frequency wordlists used for cue selection. On the other hand, in defence of Meara and Fitzpatrick’s (2000) original list of clues, it also might be argued that some imbalance in the number of responses supplied by learners should not be considered detrimental. The fact that learners may not know all of the cue words presented in Lex30 can perhaps be a good way to distinguish between lower and higher proficiency subjects. However, some care should be taken with this view as Lex30 is designed as a test of productive rather than receptive knowledge.

The same graph also sheds some light on a second important issue. It shows that there is considerable variation in the proportion of more frequently occurring words (1K words) to less frequently occurring words (1K+) produced by SA participants. Some cue words seem to attract a far higher number of lower frequency responses compared to higher frequency ones than others. Examples include *cloth* with 102 (1K) words compared to 306 (1K+) words, *fruit* (25 compared to 555 respectively) and *potato* (71 compared to 385). On the other hand, some other cue words encourage a greater number of higher frequency responses. Examples include *close* with 394 (1K) words compared to 68 (1K+) words, *experience* (220 compared to 52 respectively), *hope* (239 compared to 54), *map* (355 compared to 72) and *window* (393 compared to 58). This seeming failure of many of the cues to encourage responses which are evenly spaced along the frequency continuum may indicate that we need to consider a change the way in which the cue words are

selected. Meara and Fitzpatrick (2000), in their original paper on Lex30, stipulated that the cue words were chosen on the basis that they typically avoided eliciting a single, dominant primary response from test takers and went on to say that they should typically generate responses that were not common words – the formal criterion being that at least half of the responses should not be included in Nation’s first 1000 word list (Nation 1984). However, Figure 8.1 shows the weighting of certain responses towards the lower end of the frequency continuum in particular which would seem to refute this. In response to the inconsistencies that I have identified above there are possibly two ways in which they might be approached. They concern firstly, the use of different word frequency lists and, secondly, the use of certain different WAT norms which are used to help to decide the suitability of cue words. I will deal with each in turn.

8.5.2 Word frequency lists

Generally, frequency lists are compiled from corpora, and the corpora are compiled from texts that are selected in certain ways, so the nature of the texts determines the nature of the lists. However, in many cases the frequency profiling of a particular group of learners will vary according to distinctive learning paths that they have taken. In other words, the same frequency lists originally compiled from texts which are relevant for one group, by contrast, might not match the learning paths of another group if the same categorization of ‘infrequent’ words is used (Kremmel, 2016). This point will be examined further when we start considering the type of group used in our experiment and how this might be important in deciding what particular kind of word frequency list might be most appropriate.

With regards to the use of the Lex30 task in two of the experiments described in this thesis (see chapters 3 and 4), word frequency lists are first and foremost used to score subjects’ responses. A point is scored for each infrequent response provided, which is any word outside the first 1000 most frequent English words and up to a maximum of 120 words (30 cues x four responses= 120 words). In Meara and Fitzpatrick (2000), Nation’s (1984) word frequency list is used to both help with scoring the subject responses as well as with the selection of cue words. In a later study, however, the same researchers use the JACET 8000 (Ishikawa et al., 2003) word frequency list for the purposes of scoring, reasoning that a “more up to date set of frequency bands might improve the accuracy of the Lex30 measure” (Fitzpatrick

& Meara, 2004, p.71). However, they did not go so far as to use the new word frequency list to help with the selection of new cue words. The same 30 cue words have been retained and used in a number of subsequent studies (e.g., Fitzpatrick & Meara, 2004; Fitzpatrick & Clenton, 2010; Clenton, 2010; Fitzpatrick, 2012; Walters, 2012). We shall return to this point when we consider the other important factor used in the selection of cue words – word association norms.

Taking into account that the subjects in both the replication (chapter 3) and main longitudinal (chapter 4) experiments are LI Japanese speakers, it seems more appropriate to use frequency lists such as JACET 8000, which are designed for those learners. Clenton (2010) supports Fitzpatrick and Meara's (2004) use of the JACET 8000 word list and describes some of its advantages as follows (Clenton, 2010, p.186):

1. The original JACET8000 list was compiled in 2003. This is comparatively recent when compared to Nation's 1984 wordlist (used in Meara & Fitzpatrick, 2000). An updated version, The New JACET List of Basic Words (New JACET 8000) (JACET, 2016) has also been produced. Both lists have 8,000 words are categorized by level from 1 to 8.
2. JACET8000 is specifically designed for Japanese learners of English with words ranked according to the frequency at which Japanese learners encounter English words.
3. As well as being based on the Corpus of Contemporary American English (COCA), JACET 8000 also is also partially based on the British National Corpus (BNC). The use of a BNC component may be relevant if the wordlist is used with SA participants in the UK as is the case with the subjects in this thesis.

As I previously mentioned, Clenton (2010) advises that, when considering the use of Lex30 as a measurement tool, it is important to choose a frequency wordlist that closely matches the learning path of the group under examination. For example the use of the JACET 8000 list probably would not be appropriate for Spanish L1 EFL learners studying in the UK but would be appropriate for Japanese L1 EFL learners as is case in this thesis. Perhaps with this mind Clenton goes on to mention that one

of the advantages of the Lex30 task is its flexibility as a tool as it can be scored using a word frequency list selected to match the particular learner group in question.

8.5.3 Selection of cue words using norms databases

In addition to word frequency lists, Word Association (WA) norms were also used to select Lex30 cues. One of the inclusion criteria for selection of Lex30 cues was that they should not attract a strong primary WA response. *Cat*, for example, usually elicits *dog* to the exclusion of most other responses. If too many strong primary responses are produced they might lessen the likelihood of differentiating between test subjects. Using WA norms can help researchers avoid such problems. Over time databases of WA responses have been compiled and used in a number of fields ranging from psychology to applied linguistics and, in the case of Lex30, helping with the selection of suitable cue word. Examples of such databases include the Edinburgh Associative Thesaurus (EAT; Kiss et al., 1973), the Birkbeck word association norms (Moss & Older, 1996), the University of South Florida (USF; Nelson et al., 1998) and the English Small World of Words project (SWOW-EN; De Deyne et al., 2019). While these databases can be informative to a point, in terms of predicting responses to cues, there are a number of arguments against using native speakers' WA response data in this way for second language research.

1. Recent WA studies have found that native English speaker responses are not homogenous and vary over time (Fitzpatrick, 2007). Research has demonstrated that highly proficient Non Native Speaker (NNS) subjects are sometimes able to outperform NS ones in tests of productive vocabulary (McNamara, 1996; Meara, 1996). Schmitt (2010) warns that “unless a study uses very frequent stimulus words (which tend to have canonical responses) then native like behaviour is likely to consist of a wide range of responses” (p.253). This heterogeneity in native speakers' abilities and WA response patterns shows that studies which use such data should be treated with a degree of caution.
2. A point that is often made by researchers working in non-UK or US contexts is that normative response data that has been compiled from English L1 populations differs greatly from their NNS subjects. These differences range from demographic, with age differences between the groups being studied, to

differences in knowledge of local expressions and the influence of films, music and culture. The same researchers argue that the issue is both a cultural, linguistic and a temporal one and support the view that norms derived from the same communities and similar generations as the learners in question would increase accuracy (e.g. Higgenbotham et al., 2014; Munby, 2012; Racine et al., 2014).

3. Research shows that L2 learners' WA response patterns do not always begin to emulate native speakers' as their proficiency increases. In fact quite the opposite can occur. When Fitzpatrick (2009) conducted experiments with English L1 students learning Welsh she discovered that their L2 response patterns became similar to their L1 response profiles. Fitzpatrick and Racine (2014) replicated this experiment with Japanese EFL learners and also found similar results.

In some cases there might be some justification to compare particular learners with certain native English speaker groups. However, there is likely to be a problem with learners from some other backgrounds. With the Japanese learners of English under investigation in this thesis there is the important question of what kind of WA norms might be considered most appropriate. Given some of the drawbacks of using native speaker WA norms highlighted by Fitzpatrick (2009), Schmitt (2010) and others it might be wise to consider another solution. Citing the case of Japanese EFL students in Japan, Racine et al. (2014) make the suggestion that it is perhaps it is better to use data derived from a comparable population which is better matched to linguistic and cultural backgrounds of learners. With the Japanese SA participants taking part in the experiments described in this thesis, following such a similar path could prove promising.

8.5.5 Japanese Word Association Database of English

In light of the drawbacks to using L1 norms lists highlighted above, the use of norms lists compiled from highly proficient L2 users may have some merit. With Japanese L1 EFL learners participating in an SA programme, learner comparisons might be made using Non Native Speaker (NNS) WA normative data with responses provided by highly proficient Japanese learners of English. There still might be concerns with in group variability, particularly with differences in proficiency with individuals but these can be addressed by using precise proficiency and demographic

data. For example respondents can be categorized as ‘high-ability’ by using a narrow range of high scores on standardized proficiency tests (e.g., TOEIC scores of only 900 to 990 or CEFR levels C1 or C2). Using such checks on respondents’ eligibility may well help reduce problems of in-group variability.

Higginbotham et al. (2015) envisage the creation of a L2 learner norms database of between 5000 to 10,000 words bringing the project in line with prior WA databases such as the EAT, South Florida and SWOW-EN norms. Termed the Japanese Word Association Database of English (J-WADE) the compilation builds on the ‘Sapporo L2 English Norms’(Munby, 2014). This will involve four basic steps to be implemented over a period of at least three years. This includes the selection of high frequency cue words from recent frequency wordlists which will ensure that the majority of learner-respondents would already be familiar with, the creation of a data entry website and finally a results website by which the results can be disseminated to a wider audience. At the time of writing the project has not yet reached completion but still shows promising signs of doing so once sufficient funding has become available. Having access to such a database might make it a valuable resource in investigating Japanese learners’ vocabulary in the future.

8.5.6 Lex30: Methods of delivery

A further question relates to the way the Lex30 task is administered to participants and whether this may have any influence on the results obtained. In the past and certainly in case of the experiments carried out in this thesis, participants have generally written down their responses to written cues. Some research has indicated that there are few differences between this method and others. Clenton (2010), for example, claims that Lex30 seems to elicit spoken and written productive vocabulary knowledge in broadly the same way. On the other hand recent research may also indicate that different methods of administration can influence response behaviour (Suzuki-Parker & Higginbotham, 2019).

In chapter 2, I reviewed an article by Baba (2002) which I critically reviews Lex30 (see section 2.8.3). She discusses the standard written format of the task where subjects read cue words and write their responses and suggests that in some cases, subjects may have only provided words that they know how to write and avoided those they are simply able to verbalise. In particular she states that they may simply have lacked the orthographic knowledge to write them. Clenton (2010)

expands on this mentioning that this concern might be important for certain L1 language groups, particularly Japanese, where the orthography is different from English. To investigate this further he conducted an experiment that compared Lex30 written response scores and Lex30 spoken response scores to see if responses that the subjects produce not being influenced by a lack of orthographic knowledge. In his study 40 Japanese L1 EFL learners completed the Lex30 using the standard written format, writing up to four response words for each cue. There was then a 6-week gap between taking written task and the second task in order to allow sufficient time to forget the cues. The subjects then took the spoken format of the same Lex30 task. This involved observing the same written instructions and cues as for the standard Lex30 test with the exception that they were told to verbalise their responses. Slightly more subjects gained higher Lex30 raw scores on the written format compared to the spoken format. However, the difference between the scores was not significant.

On the other hand, a paired *t*-test shows that subjects produced a significantly higher total number of words on the spoken compared to the written format. According to Clenton (2010), this suggests that subjects may have been affected by the conditions influencing a particular test format. For example, the written version was completed within a whole class group while the spoken version had individual subjects verbalising their responses within an audio room. Subjects gained lower raw scores on the spoken format but felt confident enough to produce a higher number of higher frequency words. At the same time they may have felt more reluctant in having to respond to the spoken test format (in front of their native speaker examiner/ teacher) and may have not given themselves time to think of enough responses, possibly explaining the lower scores on the spoken Lex30 task compared to the written task.

8.6 Discussion summary and conclusion

This section has looked at a number of changes in different forms of vocabulary knowledge during SA. These include changes in the frequency of words that SA participants produce at different stages of their experience, changes in the knowledge and use of collocations, changes in orthographical knowledge and changes in the various grammatical and semantic categories that words belong to.

It shows that participants tend to produce significantly more infrequent words as well as words in total at the end of a short-term SA programme than at the beginning. We can also see that there is a period of attrition which may affect different facets of an individuals' language knowledge at varying rates. The efficacy of the measuring tool that was used to gather vocabulary samples, Lex30, was assessed and some suggestions made on ways in which it might be enhanced.

We find that with the three analyses conducted using the data that we have gathered, there is an increase in both collocational and orthographical knowledge during the SA process which is further affected by the L1 Japanese cultural and linguistic background of the participants. We can moreover acknowledge that the Wmatrix online software tool is capable of detecting changes in some of the words produced by SA participants as well as changes in the semantic groups to which they belong.

The next and final chapter lists some of the implications that the research carried out in this thesis might have for SA's major stakeholders and describes some the strengths and weaknesses of the experimental designs that were used. It will go on to look at some new directions that future studies may take before concluding with some personal comments about some of the author's motivations which originally inspired him to carry out his work.

Chapter 9 Conclusion

This chapter will consider the research described in this thesis and discuss some implications it might have for national policy makers, universities, university students and potential employers as well as reflecting on some its possible strengths and weaknesses. It will also tentatively evaluate some new directions that may build on the results that have been reported including Semantic Diversity (SemD), environmental considerations and psycholinguistics. Firstly, SemD considers not only the number of encounters with new words which enable us to learn them but also the number of different contexts in which these words occur. This may help towards a better understanding of the impact that SA may have on the acquisition of vocabulary knowledge. Secondly, particular SA environments might also assist or hinder learning experiences and research by Briggs (2015) can demonstrate how specific activities included in an SA programme, particularly out of class activities, can affect also vocabulary acquisition. Thirdly, psycholinguistic measures such as the Glasgow Norms (Scott et al., 2019) might inform us if there is any degree of emotional impact experienced by SA participants when they enter an unfamiliar SA environment. Depending on the individual involved there may be the tendency to produce words which might reveal valuable insights about changes in their emotional state during the SA experience. Finally, it will conclude with some personal comments about the motivations for carrying out this research, as well as the author's own feelings about some of the findings.

9.1 Some practical implications of this study

As previously mentioned in section 2.3, Tanaka and Manning (2018) identified a number of stakeholders involved in SA programmes. These included governments, academic institutions, parents and students and potential employers. The results of this study may help them decide that a particular SA experience will justify investment of time and financial resources. Programme cost and length together with corresponding predictions about language proficiency improvement and how far relevant skills or work experience are gained are likely to be key decision-making factors.

Governments will view shorter programmes as being able to benefit a greater number of people and thereby easier to justify to the tax paying public. As discussed in the sections 2.2.1 and 2.3, the Tobitate “Young Ambassador Programme” scheme was started with the aim of doubling the number of young Japanese studying abroad. Recent evidence shows that it continuing to provide support especially towards shorter SA programmes (MEXT, 2022). The study described in this thesis shows that it is possible to demonstrate increases in language proficiency and this can only lead to an increase in the popularity of such government sponsored short-term SA programmes.

For universities the results of this study can only encourage the growth of short-term programmes particularly as they can be incorporated more easily into existing curricula. With shorter and cheaper options it is possible that a greater range of programmes can be offered which may prove attractive to potential students. The author’s own institution, Nakamura University, now offers a number of SA options ranging from four weeks in the USA, UK and Canada to providing one-week experiences in China, Taiwan, South Korea for those students who want to learn languages other than English. Another advantage is that SA can lead to the formation of international partnerships between universities which can lead to long-lasting and mutually beneficial ties. In the case of Nakamura University short-term SA is helping to maintain relationships with the University of Hawai’i Kapi’olani Community College (USA), Meiho University (Taiwan) and Kangwon National University (South Korea). The results of this study can also have implications for teaching practice. They could help inspire an increase in those activities which focus more on functional situations and resolving communication challenges. The results may also underline the importance of adequately preparing participants for their SA experience by increasing the teaching of more specialized areas of language.

For students themselves the study might serve to increase their feelings of capacity for self-growth and awareness. The fact that a particular aspect of language proficiency (productive vocabulary knowledge) is possibly measurable can only lead to an increase in learner motivation. A short-term SA experience has been shown fit in well into a busy university schedule leaving students with sufficient time to accommodate other needs such as their studies, job hunting and part time work. For most students the SA experience has proved enjoyable, brings tangible results and contributes towards obtaining a better job on graduation. For parents, too, who are

often just seeking happiness for their children, shorter-term programmes are more affordable, provide measurable results and can offer brighter employment prospects for the future.

Lastly, the results of this study might strengthen employers' positive feelings about the potential advantages and benefits that short-term programmes may bring for future employees. As mentioned in section 2.3 companies are not always looking for language proficiency. Other skills, social and cultural awareness and communication are sometimes just as important. Many employers are keen to strengthen their overseas operations and encourage a willingness to integrate with the wider international community (Knight, 2004). Companies are likely to view students who have SA experience as potential applicants who have already demonstrated an ability to adjust to an unfamiliar environment and are capable of meeting unknown future challenges.

9.2 Strengths and weaknesses of the research

Concerning the strengths of the research carried out in this study there are perhaps four main areas which can be identified. Firstly, the data was collected at four time points not just two as is the case with many previous studies. The main advantage of this is that it enabled us to see if and to what extent predeparture activities and preparation had an effect on SA participants productive vocabulary knowledge. Using a delayed post-test also revealed if there was any decrease or attrition of knowledge once the SA experience had finished. In this way the study attempted to carry out a measurement of the long term effect of a short term programme.

Secondly, the study shows that certain aspects of language proficiency are detectable even within the short duration of a short-term SA programme. This was in contrast to earlier attempts in SA proficiency measurement (see section 2.4.2) where established generalized tests like TOEIC were not found suitable in detecting changes over short periods (Drake, 1997). Providing some means whereby participants can measure early changes in their proficiency level can only serve to increase overall levels of motivation and is likely to make such programmes more attractive to future applicants.

Thirdly, the study involved the use of only a single task, Lex30 data collection, administered on four separate occasions before and after the SA programme. This hugely simplified administration and data processing. After collection the data from the single task was then utilized in a number of further ways. Fourthly, the homogenous nature of the group might have been some advantage. Proficiency levels were difficult to judge (for associated problems see section 5.4) but the age, gender and level of SA experience might allow a degree of confidence in the data gathered.

Every research project will have some weaknesses and this study described in this thesis is no exception. Firstly, only data from a single short-term programme to UK was collected. A look at multiple programmes with a view to making some sort comparison between them over time might have been desirable. This could have revealed differences in the ability of participants to produce new vocabulary items depending on difference in programme duration and L2 environment. It is also important to note that the number of participants involved in the study was limited. A higher number would have produced more convincing results. With some analyses, most notably with collocations and orthography the paucity of data did not help with making far reaching conclusions.

Secondly, there was some uncertainty concerning SA participant proficiency levels (see section 5.4). This was important because it was created some difficulties in scrutinizing any changes in vocabulary knowledge that were taking place. It underlines the need for the use of alternative measures apart from Lex30 particularly in the early stages an SA programme. This lack of a dependable means proficiency level measurement might lead to concern with the accuracy of some analyses. This is particularly true with the collocation analysis where the means of dividing participants into different proficiency groups using Lex30 scores was perhaps oversimplistic.

Fourthly, the study was only able to collect a limited amount of data for some of the later experiments. For example, with the collocation assessment (chapter 5) test-takers were not specifically told to produce collocational responses to the Lex30 cue words. Adjusting the design of the Lex30 testing instrument so as to encourage participants to produce a greater number of words for analysis might be possible. Similarly, with the orthography experiment (chapter 6) the design could be modified so that participants would tend to produce (write down) the same words at successive

time points instead of relying on chance. This would provide a greater number of words for analysis.

Fifthly, there could be some concern with the Wmatrix analyses. The fact that test participants were limited to producing one word responses to Lex30 cues meant that the data that was produced makes comparisons with corpora difficult. Most available corpora, particularly BNC are comprised from numerous examples of extended discourse. In a future study it might make more sense to examine samples of discourse from SA participants instead of one word responses. These could make any comparison with existing corpora more meaningful.

Perhaps a final weakness with the study is that it might not take sufficient account of changing attitudes towards ‘Global Englishes’ (Galloway & Rose, 2018). On many occasions during the experiment the Japanese L1 participants produced culturally influenced vocabulary items as responses to Lex30 cues. Although a number of these were considered ‘acceptable’ and were credited as such others did not meet the set criteria. This is particularly true where collocations were considered. Some were deemed unacceptable by native speaker standards but would likely meet local non-native definitions of a typical collocation such as *curry rice* or *silver seat* (see section 5.2.2). If Dewey’s (2012) point into account which says that learners should not be penalized for producing forms of language which remain intelligible, then acceptance of more loan words and crediting non-standard collocations might bring very different results.

9.3 Semantic Diversity (SemD) and SA

A common view is that one of the ways that we learn words successfully is through the number of times we happen to encounter them. Word frequency (i.e. the number of times that a word is encountered) has been proved to be a powerful determinant of word recognition time, with high frequency words recognized more rapidly than low frequency words (e.g., Broadbent, 1967; Forster & Chambers, 1973; Krueger, 1975). However, some research suggests that the gradual accumulation of lexical knowledge is more complicated than simply counting learners’ encounters with individual words (Adelman et al., 2006). Instead, it implies that learning may be improved by meeting words in a variety of different semantic contexts so that some measure of contextual variation may provide a richer foundation for the vocabulary learning experience. One way of defining this contextual variation is through a

measure known as *semantic diversity* (Hoffman et al., 2013). The semantic diversity (sometimes abbreviated as SemD) metric is calculated using latent semantic analysis and is meant to reflect the average semantic similarity across all of the contexts in which a word occurs. To calculate SemD for a particular word, Hoffman et al. (2013) used the British National Corpus (BNC; British National Corpus Consortium, 2007) to examine all of the contexts in which the word appeared and managed to calculate SemD values for a total of 31,738 words. When the contexts were similar, this suggested that the word was associated with a limited range of meanings and tended to be unambiguous resulting in a lower SemD value. When the contexts associated with a given word appeared noticeably different to one another, this suggested that the meaning of the word was more ambiguous resulting in a higher value. Hoffman et al. (2013) give examples of two words which vary in the degree in which they are connected to a particular context: *perjury* and *predicament* (2013, p.719). *Perjury* appears in a very limited number of contexts pertaining to courtrooms and the use of legal language. The word alone gives a large amount of information about the situation in which it is used such as a witness in a court telling a lie under oath. The word *predicament*, on the other hand, describes a challenging dilemma which can occur in a wide range of different contexts such as a cat unable to climb down a tree or a politician accused of corruption. A word like *predicament* gives language learners little information about the situation in which it is used although Hoffman et al. (2013) argue that it has a “different semantic flavour” as the two contexts previously described slightly alter the way in which the word is interpreted (2013, p.720). The difference in contextual variability between the two word examples cannot be detected by normal definitions of semantic ambiguity (as both words have a single meaning) but nonetheless the different contexts in which they occur still tend to alter their meaning. SemD is used by Hoffman et al. (2013) to describe the measurement of the degree to which the different contexts in which a given word appears vary in their meaning. The contexts in which the word *perjury* occurs are likely to have a similar overall meaning, whereas the contexts in which *predicament* can be found are likely to differ substantially. Further information about calculation methods and examples can be found in Hoffman et al. (2013, pp. 721-722). Given that the SA experience gives learners an opportunity to learn words which they may have encountered before, in a range of new contexts, SemD might offer a way to quantify and better understand the impact of this.

Hsiao and Nation (2018) looked at how high frequency words are experienced in more diverse contexts over an individual's language experience. Using a similar methodology to Hoffman et al. (2013) they applied latent semantic analysis on a 35-million-word corpus of texts written for children to derive SemD values which quantified the similarity of all the contexts a word appears in. After conducting three experiments with 6–13-year-old children involving reading aloud and lexical decision making, they found that high SemD value words were responded to faster and read more accurately than low SemD value words. The authors also discovered that frequency, document count and age of acquisition were also significant predictors of reading behaviour. Using SemD values they demonstrated that contextual variability contributed toward word learning and the development of lexical quality, independently from the effects of word frequency.

Further research into children's reading was carried out by Pagán et al. (2020) which reinforced this earlier finding. Pagán et al. wanted to see how SemD could influence children's lexical decision making and reading aloud skills. Their study investigated the effects of SemD and word frequency by monitoring children's eyes movements as they read target words embedded in sentences. The reasoning was that if SemD and frequency reflect different aspects of experience that influence reading in different ways, they should show independent effects and perhaps even different processing signatures during a reading task. In their experiment 49 children (all aged nine years-old) read sentences containing various combinations of high or low frequency and high or low diversity words. Pagán et al.'s findings indicated that the nature of previous experience with words, not just the amount of experience, can shape reading development in children. They confirmed that many words with high SemD values tend to be ambiguous (or polysemous) and that SemD was capable of capturing shades of meaning based on contextual usage in a way that is continuous and graded. Their conclusion was that when a high SemD value word is experienced in a new context it is easier to identify independent of any frequency ranking it may have. Words with high SemD values were found to be easier for children to process during reading tasks independent of their frequency.

Both of these studies showed high frequency and high SemD value words being read more easily. Most importantly they demonstrated that variations in the amount and nature of contextual experience influence the degree to which words can

be processed and identified during reading independent of their frequency of occurrence.

Building on this earlier research with children's reading skills it may be possible to design an experiment to find out if the numerous and diverse nature of contexts encountered in an SA environment might actually influence the opportunity and ease with which adult SA participants can acquire new words. The range of contexts would likely range from different styles of classroom learning to travel and tourist situations or from participating in host family interactions to understanding both written and spoken signs and instructions. Such experiences might be then compared with regular Japanese L1 EFL learners in their home country typically attempting to acquire new vocabulary within the narrower confines of a language classroom.

There would be a number of challenges. The major disadvantage is that the earlier studies that I have described focus on children's reading skills and how quickly certain words could be recognized and then read aloud. For the purposes of this thesis we are more interested in focusing on adults being able to acquire productive vocabulary knowledge during a short-term SA experience. One possible approach might be the creation of two separate language corpora. The first would be an *SA participant corpus* consisting of words that are normally encountered during an SA programme. Material would be collected through transcribed classroom recordings, host family interactions, interactions with learners from other countries and a range of travel and tourist situations. The second or *Home study corpus* would incorporate teaching materials currently used during university classes using samples of textbook language and transcripts of classroom interaction. In the final step the two corpora could be analyzed to examine whether there are any differences between them particularly concerning SemD value and word frequency variables. If an SA environment can actually provide a much wider range of learning contexts which can be of benefit to acquiring vocabulary knowledge, the results might reveal that SA participants are more likely to use high SemD value words in a greater number of different ways. If such a study can show promise then this might help encourage the development of new SA programmes which more strongly emphasize a higher number of learning contexts in which to acquire new words.

9.4 Out-of-class contact: New opportunities for learning

Some research has shown how different environmental language influences can impact the process of vocabulary acquisition. Can a range of SA environments, which could include factors like travel and tourist interactions, classroom study and interactions with the host family, ultimately affect vocabulary acquisition? In this section we shall look at how the location and nature of different SA programmes can affect vocabulary-related language outcomes and how there is a potential for a comparative study using Lex30 data to reveal differences in the degree and nature of vocabulary uptake resulting from different kinds of SA programme and location.

Briggs (2015) has described some of these factors when looking at the differences in vocabulary acquisition between SA groups. In particular, she looks at the relationship between informal out-of-class language contact and vocabulary gain of SA students noting that anecdotal evidence and survey results show that an SA experience does not always result in increased exposure to L2 due to individual differences in students and their study environment. Briggs (2015) examines firstly, the types of informal language contact that SA learners identify with most and secondly, if the type of contact, SA location and length of stay has any effect on vocabulary knowledge gain. She uses the Language Contact Profile (LCP) questionnaire to measure the type and degree of out-of-class language contact, learners' personal profiles and learning environment (LCP; Freed et al., 2004a). The LCP has been widely used to provide evidence for the relationship between language contact and various forms of language gain. The results of her study showed that the language contact that learners most identified with were simple requests for information and receptive activities and that length of study and SA location sometimes had an effect on the type of informal language contact experienced. Her study concludes with a call for the inclusion in SA curricula of guidance for learners on how to plan, manage and manipulate informal language contact for maximum linguistic gain.

There is some potential for a future comparative study using some of the lessons learned from Briggs's research. One direction to take is to relate particular SA locations to vocabulary gains in general. Briggs found that it was possible to determine the types of informal language contact participants experienced when she discovered certain locations offered greater opportunities for informal social conversation due to particular host family arrangements. One further interesting

point made by the author relates to the difference of out-of-class language contact opportunities between countries (Briggs, 2015, p.138). Whether this is due to the specific country, or in the way that particular institutions provide a different variety of situations for their students to maximize their opportunities for language use, is unclear. It does mean, however, that future SA participants might want to consider a number of destinations possibly due to the greater variety of out-of-class language opportunities that they could offer. In this thesis I have used Lex30 data to reveal information about vocabulary uptake during SA, in relation to the number and frequency of words produced, spelling development, and the semantic domains of new vocabulary. Future research using a LCP-like tool to measure language contact opportunities in a range of English-speaking countries together with Lex30 to measure changes in productive vocabulary knowledge might yield some valuable and interesting results. In the case of Japanese students in the UK my own experience has shown me that students' learning environment, particularly with the wide variability in the degree of interaction and support offered by host families, can have a considerable impact on improvements in vocabulary proficiency.

9.5 The emotional impact of SA

The effect of an SA period can be personal and emotional as well as linguistic, and to date little research is available on the interaction of these two things. An exception is Tracy-Ventura et al. (2016), who investigated the emotional impact that SA might have on participants which might affect their ability both to acquire new vocabulary and to produce certain kinds of words. Using both quantitative and qualitative methods Tracy-Ventura et al. (2016) investigated personality changes affecting British students (N=58) studying in France and Spain. Similar studies in the past have found SA linguistic interactions to be beneficial for linguistic and socio-pragmatic gains, as well as increased motivation, intercultural adaptation, cross-cultural awareness, and interpersonal communication skills (e.g., Amuzie & Winke, 2009; Freed et al., 2004b).

Demonstrating changes in the SA participants' personalities could go some way towards explaining why some individuals might be able to acquire new vocabulary more easily than others although there is likely to be some difficulty in proving how far this may be the case. In Tracy-Ventura et al.'s (2016) study, a number of personality changes were measured quantitatively using an innovative

measurement tool: Multicultural Personality Questionnaire (MPQ; Van der Zee & Van Oudenhoven, 2000). The MPQ investigates factors such as cultural empathy, open-mindedness, social initiative, emotional stability and flexibility. In addition a reflective interview was conducted at the end of the SA experience and analyzed qualitatively to investigate whether students noted any personality changes. The MPQ results showed that statistically significant changes had taken place over time on the “emotional stability” factor only. This where students can demonstrate their ability to remain calm and to be able to handle difficult and stressful situations effectively. This result is supported by the reflective interviews as 77% of participants mentioned feeling more confident and independent after residence abroad. Based on these findings, the authors conclude that SA appears to be an example of a type of social investment with the potential to positively affect the emotional stability of anyone undertaking the experience as SA participants.

A second interesting example of the effect that SA’s emotional impact may have on language acquisition is related to the “Glasgow Norms”. The Glasgow Norms (Scott et al., 2018) is recent research which examines some psycho-linguistic aspects of language that SA participants, for example, might produce. A possible aim for any future research to explore is if there is any degree of emotional impact experienced by SA participants which might be measured by attaching values representing the emotional impact of words that they tend to produce. In some cases, perhaps for the very first time in their lives, SA participants are forced to be independent and to be self-sufficient and this may, perhaps, affect the kind of words that they acquire or produce. The Glasgow Norms are a set of normative ratings for 5,553 English words based on 9 psycholinguistic dimensions including arousal, valence, dominance, concreteness, imageability, familiarity, age of acquisition, semantic size, and gender association. The authors state that the norms can provide researchers ratings on specific areas of interest which can facilitate investigations into new lexical dimensions. For researchers interested in language acquisition during SA, the Glasgow Norms could provide a valuable resource helping them understand the kinds of words that learners are likely to produce and whether there are any underlying psycholinguistic reasons for them doing so.

To illustrate the way in which the Glasgow Norms might be used, Table 9.1 shows values attached to selected cue words from the Lex30 task used in this thesis.

Along the top are attached various psycholinguistic categories. For example, the first

Table 9.1

Glasgow Norms applied to selected Lex30 cue words

	Arousal	Valence	Domin	Conc	Image	Fami	Age of acquisit	Size	Gender
attack	5.09	1.94	4.15	4.36	5.05	5.63	3.53	5.15	5.50
cloth	3.56	5.48	4.97	6.39	6.48	6.10	2.72	2.88	3.50
dirty	4.91	2.97	4.26	4.64	5.44	6.27	2.06	3.97	5.03
disease	4.06	1.76	3.18	5.06	4.73	5.72	4.03	4.91	4.06
furniture	3.38	5.51	5.24	6.19	6.47	5.81	3.37	4.72	3.34
hope	7.03	8.29	6.13	4.45	3.50	6.28	3.35	5.79	2.93
rest	3.63	6.44	5.06	3.49	4.21	6.00	2.54	3.29	3.67
rice	3.57	5.20	5.09	6.74	6.45	6.18	2.27	1.66	3.61
science	6.03	7.06	5.62	4.37	4.55	6.52	3.51	6.11	5.26

three categories are *arousal*, *valence* and *dominance* which all measure different forms of a word's emotional impact, the value of which is displayed on a 9-point scale. The words *hope*, *science* and *attack* seem to indicate a high degree of **arousal** (7.03, 6.03 and 5.09 respectively) while the word *furniture* (3.38) does not. **Valence** represents the perceived value or worth of a word in positive or negative terms. *Hope* is regarded as a very positive word (8.29) while *disease* (1.76) is most certainly not. The next category **dominance** is the degree of control that the user feels they have when using the word with *hope* (6.13) being high and *disease* being low (3.18). Remaining measures at the top of the table use a 7-point scale. **Concreteness** expresses the degree to which something can be experienced by our senses so words like *furniture* and *cloth* are both concrete while *rest* is quite abstract. **Imageability** represents the degree of effort that learner might experience in generating a mental image of a word with words like *cloth* and *furniture* much easier to imagine than *hope*. Next two measures are related to word frequency. With the cue word examples given there is little variance which is not surprising as all appear within a the same frequency band of commonly occurring words most frequently occurring words. Finally, the **gender** category expresses which gender is more likely to associated with certain words. *Attack* (5.50) and *science* (5.26) seem more male oriented than words like *furniture* (3.34) and *hope* (2.93).

For the purposes of investigating how SA may affect vocabulary that participants are more likely to produce, the Glasgow Norms shows some promise. They may give some evidence about the underlying emotional state of the SA participant and whether certain individuals, who tend to produce more words belonging to a certain category, are likely to be more successful language learners.

9.6 Final thoughts and conclusion

The main motivation for starting this thesis in the first place is the difficulty I have experienced over the years with assessing the change in language proficiency of learners as a result of short-term SA programmes. Stakeholders including schools and universities, parents and the learners themselves have continued to demand evidence of language proficiency improvement. I found that when attempting to use general tests such as TOEIC and similar instruments I was unable to detect any significant changes in proficiency. At the same time I knew that the learners who had returned from their SA experiences were not the same as those who had embarked on their journeys weeks before. Most participants returned to their home country with an increased level of confidence, more motivated and more willing to tackle complex communicative tasks than before. Measuring and quantifying such changes in their behaviour was going to be challenging but I was convinced there that were ways in which some measurement of their language proficiency might be achieved.

I thought that the first important thing to consider was to concentrate on a single aspect of language knowledge in the hope of isolating and detecting more subtle changes that might be taking place. With the central role that vocabulary plays in language learning I wanted to consider this aspect in more detail. Looking across the range of different vocabulary tests and noting evidence provided by Fitzpatrick (2003), Meara (2005), Fitzpatrick and Clenton (2010), Walters (2012) and others such it became clear that it seemed that Lex30 offered a possible means of measuring short-term changes in productive vocabulary knowledge. I had noticed that returning learners seemed more willing to use their language than they were before their SA, so measuring productive vocabulary knowledge seemed to make sense.

Previous studies as well as the experiments reported in this thesis indicate that the Lex30 productive vocabulary task is capable of detecting change in vocabulary knowledge over a short space of time (e.g., Fitzpatrick & Clenton 2010; Walters 2012). The longitudinal study in chapter 4 demonstrates that a significantly

higher number of infrequent words (1K+) are produced at Time 3, after SA, than before SA at Time 2. This factor alone perhaps provides the strongest evidence that changes indeed take place during a short-term experience. While the simple distinction between low frequency (1K+) and high frequency (>1K) words is clear, a more detailed examination of words produced by SA participants using narrower 500-word bands (Kremmel 2016) did not yield such conclusive results. The Lex30 task enabled us to collect a vocabulary sample from our 38 participants at four test times. With any measurement tool, however, there is always room for improvement and it is hoped that the suggestions made in relation to this in chapter 8 will encourage others to continue to further develop Lex30. A few modifications may help this useful measurement tool gather even more data in the future.

In addition to frequency, the vocabulary samples elicited from SA participants at multiple time points were subjected to three further forms of analyses described in chapters 5, 6 and 7. The degree of change observed in each case was variable. With collocation it is evident that there is a slight increase in collocational knowledge during SA which supports findings by Alsakran (2011), Fitzpatrick (2012) and Alqarni (2017). Previous research indicates that increases tend to occur more often with lower proficiency learners so the same low proficiency level of SA participants might afford a similar potential for improvement. Two further points about the collocation study can be made. The first is that there was no specific requirement for Lex30 test takers to produce collocational responses in reply to the cue words provided which inevitably resulted in fewer being produced. A way to overcome this limitation to help encourage more collocational responses might be to adapt Lex30 in some by following recommendations given by Brown (2018). A second point is that a number of collocational responses produced by our Japanese SA participants seem culturally biased. In usual cases the methodology calls for collocations to be identified by using a specialized dictionary listing 'acceptable' L1 options. However, many words produced by Japanese L1 EFL learners do not demonstrate L1-like collocational knowledge but rather show how Lex30's L2 cues can stimulate them to produce a variety of L1 inspired cultural concepts and expressions.

With orthographical analysis it appears that some improvement in spelling accuracy is taking place perhaps indicating a gradual progression or deepening of knowledge (Zareva, 2012). As explained in chapter 6 there is a general tendency for

participants to produce similar or identical words at different time points in response to the same cue words. However, an important limitation is that participants produce such responses in only a limited number of cases. A further interesting point that was found in the study is that many misspellings are likely caused by L1 interference, especially relating to pronunciation. This is shown with our Japanese SA participants who produced a high number of misspelled words involving the well-documented confusion between *l* and *r*. Perhaps the most important discovery, however, is that evidence suggests that improving speaking skills, which are normally the main focus of many SA programmes, can in turn actually help with an improvement in spelling accuracy (Tuladhar & Akatsuka, 2017).

The use of Wmatrix also yielded some interesting results in its analysis of language produced by SA participants. The online measuring tool was used to examine keywords, that is words which appear more frequently or infrequently than expected, Parts of Speech (POS) and the semantic domains that different words belong to. It appears that SA does indeed have some impact on the language produced during the experience. For example, more words connected with semantic domains connected with an SA experience like *food*, *clothes*, *colour* and *education* were produced than would normally be expected. What is also evident is that using responses from a WAT like Lex30 will mean that many more nouns will be elicited (around 70% of the total) than words from other grammatical categories. This results in a wide area of language being overlooked, particularly in non-noun POS categories. Devising additional methods to collect language, perhaps containing natural discourse rather than individual words, from Japanese participants using similar methods as Lin (2014; 2017), would go some way toward providing a more balanced sample of language.

Another interesting feature of the study is what is revealed about cultural backgrounds. Like the British and Taiwanese participants in Lin's studies, Japanese SA participants tended to produce a number of words associated with Japanese cultural norms demonstrating that the choice of words within each semantic domain seem to demonstrate cultural and social differences. What the Wmatrix results seem to suggest that there are a number of important changes in the language produced during SA especially in the case of individual words and the semantic domains that they belong to. It is likely that this is due to participants' response to the changing SA environment. It is also evident at the same time that they are likely to choose

individual words perhaps from certain semantic domains, which more strongly reflect their own cultural background. Most importantly the study shows that L1 linguistic and cultural influences can still continue to affect the way in which learners may view a particular L2 environment, an idea which may help contribute towards a better understanding of the design of future SA programmes.

On the surface SA would seem to present a magnificent opportunity for participants to learn English across a wide variety of contexts. According to studies by Hoffman et al. (2013), Hsiao and Nation (2018), Pagán et al. (2020) this should, at the same time, afford learners some advantages when acquiring new language. Their research suggests words which are encountered more widely in a greater number of contexts are learned more easily and can be reproduced more accurately. In other words such words will be easier to identify independent of any frequency ranking it may have. The potential of Semantic Diversity (SemD), or the degree to which a word can appear over a wider number of contexts, has been tentatively explored in this thesis and suggests that some other factor apart from word frequency might be affecting the ease in which SA participants are able to produce new words. Future research into diverse learning environments and how they might affect language acquisition might further strengthen the overall appeal of SA programmes.

Not all SA experiences are the same. In some instances participants will experience an environment which is more conducive to learning new language. They might stay with a supportive host family, forge lasting friendships and be able to rely on help from their teachers and schools. Within their schedule they may be given a range of opportunities to travel, undergo new experiences and practice what they have learned in the classroom. On the other hand certain participants might encounter practical and emotional difficulties in their new environment and be offered little or no support. They might tend to only associate with others from their own L1 group or be slow to assimilate with L2 speakers. Research by Briggs (2015) shows how important some of these factors can be and how far they can affect the acquisition of a new language. The different changes in vocabulary knowledge described in this thesis will depend, in part, on these factors which will affect each individual in different ways.

The emotional impact of SA should not be underestimated. Research by Hunley (2010), Ward et al. (2001) and Furnham and Trezise (1981) has examined how participants might be impacted by the need to manage stress while coping with

an unfamiliar environment. There are also adverse effects of culture shock and a need to try and move towards positive psychological outcomes regarding SA experience. In chapter 9 I briefly discussed the Glasgow norms (Scott et al., 2019) as being one psycholinguistic measure which might inform us of the degree of emotional impact that can be experienced by SA participants when they enter an unfamiliar environment. Clearly not all SA environments and not all SA programmes are the same: in their own way each will influence the ease and degree to which participants will be able to acquire the knowledge of a new language.

As outlined in my introduction to this thesis, until now the potential advantages of short-term SA programmes have been difficult to demonstrate. Many of the benefits can only be qualitatively measured and changes in proficiency level are difficult to detect. Some changes are relatively straightforward in their measurement such as the significant increase in the number of infrequent words that learners returning from SA produce, but other changes can only be identified through more sensitive and nuanced approaches. It is hoped that some of the methodologies used in this thesis can be further refined in order to obtain more conclusive results. I have tried to firstly, show that productive vocabulary can change in a number of different ways ranging from variations in word frequency and type to shifts in semantic categorization and spelling accuracy. Secondly, I have attempted to demonstrate that such changes can take place within very short periods and finally, that environmental factors, not all of them obvious or easy to measure, can also affect the ease in which new vocabulary knowledge can be acquired. My hope is that the studies reported in this thesis might encourage a rethink of such programmes in the future by demonstrating that identifying changes in productive vocabulary knowledge is indeed possible.

References

- Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychology Science* 17 (9), 814-23.
doi: 10.1111/j.1467-9280.2006.01787.x. PMID: 16984300.
- Aizawa, K. (2006). Rethinking frequency markers For English-Japanese dictionaries. In M. Murata., K. Minamiide, Y. Tono & Ishikawa. S. (Eds.) *English Lexicography in Japan (pp. 108-119)*. Tokyo: Taishukan.
- Aizawa, K., & Iso, T. (2008). Identifying the minimum vocabulary size for academic reading *Annual Review of English Language Education in Japan (ARELE)* 19, 121–130.
- Aizawa, K., Yamazaki, A., Fuji, T., & Iino, A. (2009). The relationship between vocabulary knowledge and reading comprehension skills used on reading tests *Annual review of English Language Education in Japan (ARELE)* 20, 111–120.
- Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and second language: Processes and development*. Basingstoke: Palgrave Macmillan.
- Alderson, J. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.
- Alejo González, R., & Piquer Piriz, A. (2016). Measuring the productive vocabulary of secondary school CLIL students: Is Lex30 a valid test for low-level school learners? *Vigo International Journal of Applied Linguistics* 13, 31-53.
- Alenazi, O. (2018). Spelling difficulties faced by Arab learners of English as a foreign language. *Arab World English Journal (AWEJ)* 9 (2), 118-126.
- Alqarni, I. (2017). The impact of length of study abroad on collocational knowledge: The case of Saudi students in Australia. *Advances in Language and Literary studies* 8 (2), 237-242.
- Alsakran R. (2011). *The productive and receptive knowledge of collocations by advanced Arabic-speaking ESL/EFL learners*. [Unpublished doctoral dissertation]. Colorado State University, USA.
- Amed, J. (2011). *Acculturation and academic phenomenon of Saudi Arabian scholarship students entering medical and health education* [Unpublished doctoral dissertation]. Faculty of Medicine, University of Sydney, Australia.
- Ammade, S., Ramadhani, M. F., & Rahman, A. W. (2023). Google Translate as English Learning Tool Assistance for non-English Departments Students: Students' perception. *Klasikal : Journal Of Education, Language Teaching And Science*, 5 (1), 167–181. <https://doi.org/10.52208/klasikal.v5i1.600>

- Amuzie, G. L., & Winke, P. (2009). Changes in language learning beliefs as a result of study abroad. *System*, 37 (3), 366-379.
- Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper 16*, pp. 22-31. UCREL, Lancaster University.
- Arnaud, P., & Savignon, S. (1997). Rare words, complex lexical units and the advanced learner. In J. Coady and T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp.157-173). Cambridge University Press.
- Asaoka, T., & Yano, J. (2009). The contribution of study abroad programs to Japanese internationalization, *Journal of Studies in International Education* 13 (2), 174-188.
- Avello, P., & Lara, A. (2014). Phonological development in L2 speech production during study abroad programmes differing in length of stay. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp.137-165). Amsterdam: John Benjamins.
- Baba, K. (2002). Lex30: Review, *Language Testing Update*, 32, 68-71.
- Bachman, L. (1990). *Fundamental considerations in Language testing*, Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*, Oxford: Oxford University Press
- Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System* 21 (1), 101-114.
- Bahrack, H. (1984). Fifty years of second language attrition: Implications for programmatic research. *Modern Language Journal*, 68, 105-118.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker-Smemoe, W., Dewey, D., Bown, J., & Martinsen, R. (2014). Variables affecting L2 gains during study abroad. *Foreign Language Annals* 47 (3), 464-486.
- Bardovi-Harlig, K., & Bastos, M. (2011). Proficiency, length of stay, and intensity of interaction and the acquisition of conventional expressions in L2 pragmatics. *Intercultural Pragmatics* 8, 347-384.
- Barfield, A. (2009). Following individuals' L2 collocation development over time. In A. Barfield and H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 208-223). Basingstoke: Palgrave Macmillan.

- Barfield, A., & Gyllstad, H. (2009). Introduction: Researching L2 collocation knowledge and development. In A. Barfield and H. Gyllstad (Eds.), *Researching collocations in another language: Multiple Interpretations* (pp.1-18). Basingstoke: Palgrave Macmillan.
- Basetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics*, 36 (1), 67-91.
- Barquin, E. (2012). *Writing Development in a Study Abroad Context* [Unpublished doctoral dissertation]. Universitat Pompeu Fabra, Barcelona.
- Bauer, L., Nation, I.S.P. (1993). Word families. *International Journal of Lexicography* 6, 253-279.
- Beaton, A., Gruneberg, M., and Ellis, N. (1995). Retention of foreign vocabulary learned using the keyword method: a ten-year follow-up. *Second Language Research* 11(2) 112–120. doi.org/10.1177/026765839501100203
- Bebout, L. (1985). An error analysis of misspellings made by learners of English as a first and as a second language *Journal of Psycholinguistic Research* 14 (6), 569-593. https://doi.org/10.1007/BF01067386
- Bell, H. (2009). The messy little details: a longitudinal case study of the emerging lexicon. In *Lexical Processing in Second Language Learners* (eds.) T. Fitzpatrick and A. Barfield (pp.111–27). Bristol, UK: Multilingual Matters.
- Benson, P., Barkhuizen, G., Bodycott, P., & Brown, J. (2013). *Second Language Identity in Narratives of Study Abroad*. NY: Palgrave Macmillan.
- Benson, M., Benson, E., & Ilson, R. (2009). *The BBI Combinatory Dictionary of English* (Third ed.). Amsterdam: John Benjamins Publishing Company.
- Berwick, R., & Ross, S.(1989). Motivation after matriculation: Are Japanese learners of English still alive after exam hell? *JALT Journal* 11 (2), 193–210.
- Bianchi, F. (2017). The social tricks of advertising. Discourse strategies of English-speaking tour operators on Facebook. *Iperstoria* 10, 3-32.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
- Bonk, W. J. (2000). *Testing ESL learners' knowledge of collocations*. Educational Resources Information Center. Retrieved 4 July, 2022, from <https://files.eric.ed.gov/fulltext/ED442309.pdf>

- Breeze, R. (2019). Emotion in politics: Affective-discursive practices in UKIP and Labour. *Discourse and Society* 30 (1) 24-43.
- Brick, B., & Cervi-Wilson, T. (2015). Technological diversity: a case study into language learners' mobile technology use inside and outside the classroom. In K. Borthwick, E. Corradini, & A. Dickens (Eds.), *10 years of the LLAS elearning symposium: Case studies in good practice* (pp. 21-30). Dublin: Research-publishing.net. doi:10.14705/rpnet.2015.000264
- Briggs, J.G. (2015). Out-of-class language contact and vocabulary gain in a study abroad context. *System* 53, 129-140.
- British National Corpus Consortium (2007). *British National Corpus version 3* (BNC XMLth ed.). Oxford, U.K. Oxford University Computing Services.
- Broadbent, D. (1967). Word-frequency effect and response bias. *Psychological Review*, 74, 1-15. doi:10.1037/h0024206
- Brown, D. (2014). Knowledge of Collocations. In J. Milton and T. Fitzpatrick (Eds.) *Dimensions of Vocabulary Knowledge* (pp.123-139). Basingstoke: Palgrave Macmillan.
- Brown, D. (2018). *Developing a measure of L2 learners' productive knowledge of English collocations*. [Unpublished doctoral dissertation]. Cardiff University, UK.
- Butler, Y., and Iino, M. (2005). Current Japanese reforms In English language education: the 2003 action plan *Language Policy* 4, (1), 25–45.
- Byram, M. (1997). *Teaching and assessing Intercultural communicative competence*. Clevedon UK: Multilingual Matters.
- Byram, M., Gribkova, B., and Starkey, H. (2002). *Developing the intercultural dimension in language teaching: a practical introduction for teachers*. Strasbourg: Council of Europe.
- Caton, T. (2013, 28-30 January). *The unexpected benefits of short Study Abroad (SA) programmes* [Conference presentation]. International conference on enhancing employability through proficiency in Indian and foreign languages, Modern College of Arts, Science and Commerce, Shivajinagar, Pune, India.
- Cervetti, G., Tilson, J., Castek, J., Bravo, M., & Trainin, G. (2012). Assessing multiple dimensions of vocabulary knowledge. *Journal of Education* 192, 49-61. DOI:10.1177/0022057412192002-308
- Chikamatsu, N. (1996). The effects of L1 orthography on L2 word recognition. *Studies in Second Language Acquisition* 18 (4), 403-432.
- Churchill, E. (2008). A dynamic systems account of learning a word: from ecology to form relations. *Applied Linguistics* 29 (3), 339–358.

- Clenton, J. (2005). Why Lex30 may not be an improved method of assessing productive vocabulary in an L2. *Studies in Language and Culture* 31, 47-59. Osaka University, Japan.
- Clenton, J. (2006). Addressing the construct validity and the reliability of Lex30 as a measure of productive vocabulary for non-native learners of English. *Osaka University graduate school of language and culture project*, pp. 1-20.
- Clenton, J. (2008). Addressing Baba: determining whether Lex30 is a reliable and valid testing measure. *Studies in Language and Culture* 34, 157-168.
- Clenton, J. (2010). *Investigating the construct of productive vocabulary knowledge with Lex30*. [Unpublished doctoral dissertation]. Swansea University, UK.
- Cobb, T. (n.d.) *Compleat lexical tutor* (Version 8.5) [Computer Software] Accessed 15 Sept 2020 at <https://www.lextutor.ca>
- Cochrane, R. (1980). The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language. *Journal of Multilingual and Multicultural Development* 1 (4), 331-360.
- Collentine, J., & Freed, B. (2004). Learning context and its effects on second language acquisition: Introduction. *Studies in Second Language Acquisition* 26 (2), 153–171.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition* 26, 227–248.
- Collentine, J. (2009). Study abroad research: findings, implications and future directions. In C. Doughty and M. Long (Eds.) *Handbook of Language Teaching* (pp.218–233). Malden, MA: Blackwell.
- Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics* 32, 45-61. doi: 10.1017/S0267190512000074
- Cook, V. (1997). L2 users and English spelling. *Journal of Multilingual and Multicultural Development* 18 (6), 474-488.
- Cook, V. (1999). *Teaching L2 spelling* [unpublished document] Retrieved on 20 September 2020 at <http://www.viviancook.uk/Writings/Papers/TeachingSpelling.htm>
- Cook, V.J. (2001). *Second language learning and teaching*. (3rd ed). London: Arnold.
- Cowie, A.P. (1998). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press.
- Crowther, J., Dignen, S., & Lea, D. (2002). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press

- Crystal, D., (1992) *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.
- Dale, E. (1965). Vocabulary measurement: techniques and major findings. *Elementary English* 42, 395-401.
- Dann, G. (1996). *The language of tourism: a sociolinguistic perspective*. Wallingford: CAB International.
- Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. London: Routledge.
- De Deyne, S., Navarro, D., Perfors, A., Brysbaert, M., & Storms, G. (2019, August 20). The Small World of Words: English word association norms for over 12,000 cue words. <https://doi.org/10.3758/s13428-018-1115-7>
- DeKeyser, R. (1990). Towards a valid measurement of monitored knowledge. *Language Testing*, 7, (2), 147–157. <https://doi.org/10.1177/026553229000700202>
- DeKeyser, R. (2014). Research on language development during study abroad. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction* Contexts (pp.313-326). AILA Applied Linguistics Series 13.
- De Vattel, E. (1844). *The law of nations, or the principles of law of nature applied to the conduct and the affairs of nations and sovereigns with three early essays on the origin and nature of natural law and on luxury* (PDF). Philadelphia: T. and J. W. Johnson – via Library of Congress. Retrieved on 1 January 2020 from http://www.loc.gov/rr/frd/Military_Law/Lieber_Collection/pdf/DeVattel_La_wOfNations.pdf
- Dewey, D. (2008). Japanese vocabulary acquisition by learners in three contexts *Frontiers: The Interdisciplinary Journal of Study Abroad* 15, 127–148.
- Dewey, M. (2012). Towards a post-normative approach: learning the pedagogy of ELF. *Journal of English as a Lingua Franca* 1 (1), 141–70.
- Dickinson, M. (2005). *Error detection and correction in annotated corpora*. [Unpublished doctoral dissertation]. Ohio State University, USA.
- Dougill, J. (2008). Japan and English as an alien language. *English Today* 24 (1), 18–22. Doi:10.1017/s0266078408000059.
- Drake, D. (1997). Integrating study abroad students into the university community. *The Language Teacher* 21 (1), 7-13.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1), 61–74.

- Duperron, L., & Overstreet, M. (2009). Preparedness for study abroad: comparing the linguistic outcomes of a short-term Spanish program by third, fourth and sixth semester L2 learners. *Frontiers: The Interdisciplinary Journal of Study Abroad* 18, 157–179.
- Ebeid, E.H., (2016, 24 May). Are loanwords a threat to the Japanese language? *Nippon.com* Retrieved on 12 September, 2020 from <https://www.nippon.com/en/column/g00195/>
- Ecke, P., & Hall, J.J, (2012). Tracking tip-of-the-tongue states in a multilingual speaker: evidence of attrition or instability in lexical systems? *International Journal of Bilingualism* 17 (6), 734–751.
- Ellis, R. (1999). Item versus system learning: explaining free variation. *Applied Linguistics* 20, 460–80.
- Fantini, A.E. (2000). A central concern: developing intercultural competence. In A.E. Fantini (ed.) *SIT Occasional Papers* (pp. 25–42). Brattleboro, Vermont: SIT.
- Fantini, A.E. (2012). Multiple strategies for the assessment of intercultural communicative competence. In J. Jackson (ed.) *Routledge Handbook of Language and Intercultural Communication* (pp. 390–405). London: Routledge.
- Farghal, M., & Obiedat, H. (1995). Collocations: a neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching* 33 (4), 315-332.
- Fielding, K., Head, B., Laffan, W., Western, M., and Hoegh-Guldberg, O. (2012). Australian politicians' beliefs about climate change: political partisanship and political ideology. *Environmental Politics* 21 (5), 712–733. doi:10.1080/09644016.2012.698887
- Firth, J. (1951). Modes of meaning. *Essays and studies*, 118-149. (Reprinted in J.R. Firth (1957). *Papers in Linguistics: 1934-1951* (pp. 1190-1215). London: Oxford.
- Fitzpatrick, T. (2003). *Eliciting and measuring productive vocabulary using word association techniques and frequency bands* [Unpublished doctoral dissertation]. Swansea University, UK.
- Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In Daller, H., Milton, J., and Treffers-Daller, J. (Eds.). *Modelling and Assessing Vocabulary Knowledge* (pp. 116-132). Cambridge University Press (CUP).
- Fitzpatrick, T. (2009). Word association profiles in a first and second language: puzzles and problems. In T. Fitzpatrick and A. Barfield (Eds.). *Lexical Processing in Second Language Learners*. Multilingual Matters, pp. 38–52.

- Fitzpatrick, T. (2012). Tracking the changes: vocabulary acquisition in the study abroad context. *The Language Learning Journal* 40 (1), 81-98.
- Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: assessing the performance of a test of productive vocabulary. *Language Testing* 27 (4), 537–554.
- Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly* 51 (4), 844-867.
- Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics* 1, 55–74.
- Fitzpatrick, T., & Racine, J. P. (2014). *Using learners' L1 word association profiles as an alternative to native speaker norms*. [Conference presentation]. AILA World Congress, Brisbane, Australia.
- Forster, K., & Chambers, F. (1973) Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627-635.
- Foster, P. (2009). Lexical diversity and native-like selection: the bonus of studying abroad. In B. Richards, M. Daller, D. Malvern, P. Meara, J. Milton and J. Treffers-Daller (Eds.). *Vocabulary Studies in First and Second Language Acquisition* (pp. 91–106). Hampshire: Palgrave Macmillan.
- Frame, A., & Brachotte, G. (2016). *Citizen Participation and Political Communication in a Digital World*. London: Routledge.
- Freed, B. (1995). What makes us think that students to study abroad become fluent? In B. Freed (Ed.), *Second Language Acquisition in a Study Abroad Context* (pp. 123–148). Amsterdam: John Benjamins.
- Freed, B., So, S., & Lazar, N. A. (2003). Language learning abroad: How do gains in written fluency compare with gains in oral fluency in French as a second language? *ADFL Bulletin*, 34(3), 34–40.
- Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004a). The language contact profile. *Studies in Second Language Acquisition*, 26 (2), 349-356.
- Freed, B., Segalowitz, N., & Dewey, D. (2004b). Context of learning and second language fluency in French: Comparing regular classroom, study abroad and intensive domestic immersion programs. *Studies in Second Language Acquisition* 26, 275–301.
- Fujita, R., Yokouchi, Y., Matsuoka, D., Nakamura, K., & Hirai, A. (2016). Examination of the Grade 2 EIKEN Test Items: Focusing on CEFR Levels, Achievement Goals, and Washback Effects. *KATE Journal*, 30, 85–97. https://doi.org/10.20806/katejournal.30.0_85
- Furnham, A. and Tresize, L. (1981). The mental health of foreign students. *Social Science and Medicine*, 17, 365–370.

- Galloway, N. and Rose, H. (2018). Incorporating global Englishes into the ELT classroom. *ELT Journal* 72 (1), 3–14.
- Gass, S., & Selinker, L. (2001). *Second language acquisition: An introductory course* (2nd Edition). Mahwah, NJ: Lawrence Erlbaum.
- Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research* 42, 237-288.
- Gobert, M. (2007). Collocational Knowledge of Gulf Students. In Davidson, P., Coombe, C., Lloyd, D. and Palfreyman, D. (Eds.) *Teaching and Learning Vocabulary in Another Language* (pp. 49-60). Dubai: TESOL Arabia.
- Gotti, M. (2006). The language of tourism as specialized discourse. In O. Palusci and S. Francesconi (Eds.) *Translating Tourism: Linguistic/Cultural Representations* (pp. 15-34). Trento: Università degli Studi di Trento.
- Gretzel, U., & Yoo, K.H. (2014). Premises and promises of social media marketing in tourism. In S. McCabe (Ed.) *The Routledge Handbook of Tourism Marketing*. (pp.491-504). London: Routledge.
- Grey, S. (2018). Quantitative Approaches for Study Abroad Research. In C. Sanz & A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (pp. 48-57). New York, NY: Routledge.
- Gu, Y. (2003). Vocabulary learning in second language: person, task, context and strategies. *Electronic Journal TESL-EJ* 7 (2), 1-26. <https://tesl-ej.org/ej26/a4.html>
- Gunion, R. (2012). What are the types and proportions of major spelling errors made by short-stay Japanese University students enrolled full-time at Newcastle University? *ARECLS* 9, 15-41.
https://research.ncl.ac.uk/media/sites/researchwebsites/arecls/gunion_vol9.pdf
- Hansen, L., Umeda, Y., and McKinney, M. (2002). Savings in the relearning of second language vocabulary: The effects of time and proficiency. *Language Learning* 52 (4), 653-678.
- Harvey, K. (2012). Disclosures of depression: Using corpus linguistics methods to interrogate young people's online health concerns. *International Journal of Corpus Linguistics* 17 (3), 349–379.
- Harvey, K. (2013). *Investigating Adolescent Health Communication: A Corpus Linguistics Approach*. London and New York: Bloomsbury.
- Harwell, M., Rubinstein, E., Hayes, W., & Olds, C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics* 17, 315-339.

- Hayes-Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. *Language Teaching* 54 (3), 297-326. doi.org/10.1017/S0261444820000658
- Heatley, A., Nation, I.S.P., & Coxhead, A. (2003). *Range* [Computer Software] Available from: <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs>
- Henriksen, B. (2012). Researching L2 learners' collocational competence and development – a progress report. *EUROSLA Monographs Series 2*, 29–56.
- Higginbotham, G., Munby, I., & Racine, J. (2015). A Japanese Word Association Database of English. *Vocabulary Learning and Instruction* 4 (2), 1-20.
- Higuera García, M. (2017). Pedagogical principles for the teaching of collocations in the foreign language classroom. In S. Torner Castels and E. Bernal (Eds.), *Collocations and other lexical combinations in Spanish: Theoretical, Lexicographical and Applied Perspectives* (pp.250–265). New York NY: Routledge.
- Hill, Y. Z. (2010). *Validation of the STEP EIKEN test for college admission* (Publication No.3429733). [Doctoral dissertation, University of Hawai'i at Manoa] Available from ProQuest Dissertations & Theses Global.
- Hirai, M. (2014, November 5). Understanding Japanese motivations for studying abroad (or not) *Recruiting Intelligence*. Retrieved on 1 January 2020 from <https://services.intead.com/blog/understanding-japanese-motivations-for-studying-abroad-or-not>
- Hoffman, P., Lambon Ralph, M., & Rogers, T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45 (3), 718–730. doi://doi.org/10.3758/s13428-012-0278-x
- Horn, E., & Ashbaugh, E. (1920). *Lippincott's Horn-Ashbaugh speller for grades one to eight*. Lippincott. <http://www.gutenberg.org/etext/33826>
- Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review* 56 (2), 309-330.
- Hoshino, Y. (2020). Vocabulary range and characteristics of words appearing in elementary school English textbooks in Japan. *Journal of the National Association for English Language Education* 31, 49-63.
- Howarth, P. (1998a). Phraseology and second language proficiency. *Applied Linguistics* 19 (1), 24-44. doi: 10.1093/applin/19.1.24
- Howarth, P. (1998b). The Phraseology in Learners' Academic Writing. In A.P. Cowies (Ed.) *Phraseology* (pp.24-44). Oxford: Oxford University Press.

- Hsiao, Y., and Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language* 103, 114–126. doi:10.1016/j.jml.2018.08.005
- Hunley, H. (2010). Students' functioning while studying abroad: The impact of psychological distress and loneliness. *International Journal of Intercultural Relations* 34 (4), 386-392.
- HSBC Report (2017). The value of education: higher and higher (pp.1-38). Retrieved 1 January 2020 from: www.hsbc.com/~media/hsbc-com/newsroomassets/2017/pdfs/171205-the-value-of-education.pdf
- Ibrahim, M. (1978). Patterns in spelling errors. *English Language Teaching* 32 (3), 207-212.
- IELTS (2020). IELTS Indicator. <https://www.ieltsindicator.com/>
- Ife, A., Vives, B., & Meara, P. (2000). The impact of study abroad on the vocabulary development of different proficiency groups. *Spanish Applied Linguistics*, 4, 55-84.
- Imao, Y. (2019). Casual Conc (Version 2.1.1) [Computer Software] Osaka, Japan. Available from: <https://sites.google.com/site/casualconc/Home>
- In'nami, Y., & Koizumi, R. (2017). Using EIKEN, TOEFL, and TOEIC to Award EFL Course Credits in Japanese Universities. *Language Assessment Quarterly*, 14, 3 274–293. <https://doi.org/10.1080/15434303.2016.126237>
- Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., Tono, Y., & Murata, M. (2003). *JACET 8000: JACET list of 8000 basic words*. Tokyo Japan: JACET.
- Ishikawa, S. (2016) *The New List of Basic Words (New JACET8000) (JACET, 2016)*. Available from <http://language.sakura.ne.jp/s/voc.html>
- JACET (Japan Association of College English Teachers) (2003). *JACET List of 8000 Basic Words*. Tokyo: JACET.
- JANU (Japan Association of National Universities) (2016). Tobitate! (Leap for Tomorrow) Study Abroad Initiative [Summary of MEXT report on Tobitate] Retrieved on 1 January 2019 from <https://www.janu.jp/eng/global-engagement/tobitate/>
- JAOS (Japan Association of Overseas Study) (2017). *Survey on the Number of Japanese Studying Abroad Report*. Retrieved on 10 January 2020 from: <http://www.jaos.or.jp/top-eng>

- JAOS (Japan Association of Overseas Study) (2021). *Statistical Report on Japanese Studying abroad*. Retrieved 8 January 2020 from https://www.jaos.or.jp/wp-content/uploads/2021/06/JAOS-Survey2021_210611_en.pdf
- JASSO (Japan Student Service Organisation) (2016). *Heisei 26 nendo kyouteitou nimotodoku nihonjingakusei ryuugakujoukyou ghousa* [The summary of results on status survey of Japanese students studying abroad 2014] Retrieved on 5 January 2020 from http://www.jasso.go.jp/about/statistics/intl_student_s/2015/index.html
- JPSS (Japan Study Support) (n.d). *Information for international students*. Retrieved 8 January 2020 from <https://www.jpss.jp/ja/univ/english/>
- Jiménez-Jiménez, A. (2010). A Comparative Study on Second Language Vocabulary Development: Study Abroad vs Classroom Settings. *Frontiers: The Interdisciplinary Journal of Study Abroad*, 19, 105-123.
- Jiménez-Catalán, R. M., & Moreno Espinosa, S. (2005). Using Lex30 to Measure the L2 Productive Vocabulary of Spanish Primary Learners of EFL. *VIAL Vigo International Journal of Applied Linguistics* 2, 27-44.
- Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *PASAA* 57, January-June 2019. Retrieved on 27th January 2020 from: <https://files.eric.ed.gov/fulltext/EJ1224421.pdf>
- Kim, H.S., Lawrence, J.H. (2021). Who Studies Abroad? Understanding the Impact of Intent on Participation. *Res High Educ* 62, 1039–1085. <https://doi.org/10.1007/s11162-021-09629-9>
- Kindaichi, H., Shibata, T., and Hayashi, N. (1988). *日本語百科大事典* [Japanese encyclopedia]. Taishukan Shoten Tokyo.
- Kinginger (2009). *Language learning and study abroad: A Critical Reading of Research*. UK: Palgrave Macmillan.
- Kinginger, C. (2011). Enhancing language learning in study abroad. *Annual Review of Applied Linguistics* 31, 58-73. doi:10.1017/S0267190511000031
- Kinginger, C. (2013). *Social and cultural aspects of language learning in study abroad*. Amsterdam: John Benjamins BV.
- Kirkpatrick, A. (2007). Setting attainable and appropriate English language targets in multilingual settings: A case for Hong Kong. *International Journal of Applied Linguistics* 17 (3) 376–91.
- Kiss, G.R., Armstrong, C., & Milroy, R. (1973). *An Associative Thesaurus of English*. EP Microfilms, Wakefield.

- Knight, J. (2004). Internationalization remodeled: Definition, approaches and rationale. *Journal of Studies in International Education* 8 (5), 5-31.
- Knight, D., Adolphs, S., and R. Carter, R. (2014). CANELC: constructing an e-language corpus, *Corpora* 9 (1), 29–56.
- Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly* 50 (4), 976-987.
- Krueger, L. (1975). Familiarity effects in visual information processing. *Psychological Bulletin* 82, 949-974. doi:10.1037/0033-2909.82.6.949
- Kuno, Y. (2014). *Shinsotsukara Kaigai de Hatarakou* [Let's Start Working Overseas right after your Graduation] .Tokyo: TCG Publications Ltd.
- Kuwabara, H. (2021, June 23). Private-sector English tests not likely for college entrance exams. *Asahi Shimbun*
<https://www.asahi.com/ajw/articles/14379298>
- Lara, A. (2014). Complexity, accuracy and fluency development through study abroad programmes varying in duration [Unpublished doctoral dissertation], Universitat Pompeu Fabra, Spain. Retrieved from www.tdx.cat/handle/10803/284229, accessed January 1 2022
- Laufer, B. (1989). What percentage of lexis is essential for comprehension? In Lauren, C., and Nordman, M. (Eds.), *From Humans Thinking to Thinking Machines* (pp.316–323). Clevedon: Multilingual Matters.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P.Arnaud and H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126-132).Palgrave Macmillan, London.
- Laufer, B., & Nation, I.S.P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16, 307-322.
<http://dx.doi.org/10.1093/applin/16.3.307>
- Laufer, B., & Nation, I.S.P. (1999). A Vocabulary-Size Test of Controlled Productive Ability. *Language Testing* 16, 33-51.
- Laufer, B., & Paribakht, T. (1998). The relationship between passive and Active vocabularies: Effects of language learning context. *Language learning*, 48 (3), 365-391.
- Laufer, B., & Sim, D. (1985). Measuring and explaining the threshold needed for English for academic purposes texts. *Foreign Language Annals* 18, 405–413.
- Leonard, K., & Shea, C. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, 101 (1), 179–193.
<https://doi.org/10.1111/modl.12382>

- Liaw, M.L. (2006). E-learning and the Development of Intercultural Competence. *Language Learning and Technology* 10, (3) 49–64.
- Lin, Y. L. (2013). Discourse functions of recurrent multi-word sequences in online and face-to-face intercultural communication. In J.Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics* (pp. 105–129). London: Springer.
- Lin, Y.L. (2014). Using key part-of-speech analysis to examine spoken discourse by Taiwanese EFL learners. *ReCALL* 27 (3), 304–320.
doi:10.1017/S0958344014000408 Published online 30 December 2014
- Lin, Y.L. (2017). Keywords, semantic domains and intercultural competence in the British and Taiwanese teenage Intercultural Communication Corpus. *Corpora* 12 (2), 279–305. DOI: 10.3366/cor.2017.0119
- Lix, L.M., Keselman, J.C., & H.J. Keselman (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research* 66, 579-619.
- Llanes, À., and Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System* 37 (3), 353-365.
- Llanes, A. (2011) The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism* 8 (3), 189–215.
- Llanes, À. (2010). *Children and adults learning English in a study abroad context* [Unpublished doctoral dissertation]. Universitat de Barcelona, Spain.
- Macintosh, C., Francis, B., & Poole, R. (2009). *The Oxford Collocations Dictionary for students of English*. Oxford University Press.
- Marijuan, S., & Sanz, C. (2017). Technology-assisted L2 research in immersive contexts abroad. In P. Costa, H. Rawal and I. Zaykovskaya (Eds.), *System* 71, 22-34. <https://doi.org/10.1016/j.system.2017.09.017>
- Mark, R. (1998). Spelling accuracy in Japanese EFL students: Some practical and theoretical implications. *Studies in Humanities, Culture and Communication* 32, 133-142.
- Martínez Adrián, M., and Gallardo del Puerto, F. (2017). The effects of language typology on L2 lexical availability and spelling accuracy. *International Journal of English Studies* 17 (2), 63-79. DOI: 10.6018/ijes/2017/2/256411
- Matsumoto, M. (2012). Effectiveness of short-term overseas English study abroad programs. *VISIO* 42, 1-10.
- Maximo, R. (2000). Effects if rote, context, keyword, and context/keyword method on retention of vocabulary in EFL classroom. *Language Learning* 50 (2), 385-412.

- McCarthy, M., & Handford, M. (2004). Invisible to us: A preliminary corpus-based study of spoken business English. In U. Connor and T.A.Upton (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics* (pp. 167–201). Amsterdam: Benjamins.
- McCrostie, J. (August 9, 2017). More Japanese may be studying abroad, but not for long. *The Japan Times*. Retrieved on 1st January 2020 from <https://www.japantimes.co.jp/community/2017/08/09/issues/japanese-may-studying-abroad-not-long/#.Wa37u00w-Ag>
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies. An Advanced Resource Book*. London, New York: Routledge.
- McManus, K., Mitchell, R., & Tracy-Ventura, N. (2020). A longitudinal study of advanced learners' linguistic development before, during and after study abroad. *Applied Linguistics* 42 (1), 136-163. <https://doi.org/10.1093/applin/amaa003>
- McNamara, T. (1996). *Measuring second language performance*. Addison Wesley Longman
- Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and Competence in Second Language Acquisition* (pp. 35– 53). Cambridge: Cambridge University Press.
- Meara, P. (1999). Self-organization in bilingual lexicons. In *Language and Thought in Development*. In P. Broeder and J. Murre (Eds.), (pp. 127-144). Tubingen: Narr.
- Meara, P. (2005). Reactivating a dormant vocabulary. In S. Foster-Cohen, M. del Pilar García Mayo & J. Cenoz (Eds.), *EUROSLA Yearbook 5* (pp. 269–280). Amsterdam: Philadelphia John Benjamins.
- Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing* 4 (2), 142-151.
- Meara, P., & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System* 28, 19-30.
- Meara, P., & Jones, G. (1990). *Eurocentres Vocabulary Size Test* (version E1.1/K109). Eurocentres Learning Service: Zurich.
- Meara, P., & Milton, J.L. (2002). *X Lex: The Swansea Vocabulary Levels Test*. Newbury, UK: Express Publishing.
- MEXT (2018). *Kotogakko gakushu shido yoryo kaisetsu gaikokugo hen eigo hen* (Explanation of the high school course: foreign languages English). https://www.mext.go.jp/content/1407073_09_1_2.pdf
- MEXT (2022) website: <https://tobitate.mext.go.jp/about/english.html>

- Milton, J. (2013) Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist and B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57–78). Amsterdam, Netherlands: Eurosla.
- Milton, J., & Alexiou, T. (2009). Vocabulary size and the common European framework of reference in languages. In B. Richards, H. Daller, D. Malvern, P. Meara, J. Milton and J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition* (pp.194-211). Palgrave: Macmillan.
- Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics* 107/108, 17-34.
- Miura, K. (2020). Developing semantic-based DDL based on a comparative study of the verb use of British and Japanese students. *Learner Corpus Studies in Asia and the World Vol 4*, 41-66. (Papers from LCSAW2019) ISSN: 2435-2632 Published by School of Language and Communication, Kobe University, Japan.
- Miyahara, F., Namoto, M., Yamanaka, H., Murakami, R., Kinoshita, M., & Yamamoto, H. (1997). *Konamamade Yoinoka Daigaku Eigo Kyouiku* [The current concerns on English language teaching at colleges and universities]. Tokyo: Shouhakusha.
- Moffitt, B. (2016). *The Global Rise of Populism: Performance, Political Style and Representation*. Stanford, CA: Stanford University Press.
- Mohle, D. & Raupach, M. (1987). The representation problem in interlanguage theory. In W. Lorsch & R. Schulze (Eds.), *Perspectives on language in performance* (pp. 1158-1173). Tiibingen: Gunter Narr.
- Moss, H., & Older, L. (1996). *Birkbeck Word Association Norms*. Hove: Psychology Press.
- Munby, I. (2014). *Sapporo word association norms lists*. Retrieved from <http://sapporowordassociationnormslists.wordpress.com/>
- Nakano, T., & Koyama, Y. (2005). e-Learning materials development based on abstract analysis using web tools. *Knowledge-based Intelligent Information and Engineering Systems (Pt 1) Proceedings, Part 1, LNCS 3681* (pp.794-800). DOI 10.1007/11552413_113
- Nation, I. S. P. (Ed.) (1984). Vocabulary Lists: words, affixes and stems. *English Language Institute Victoria University of Wellington Occasional Paper, 12*.
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524759

- NIPSSR (National Institute of Population and Social Security Research) (n.d). Retrieved January 8, 2020 from <http://www.ipss.go.jp/index-e.asp>
- Nelson, D., C.L.McEvoy, C., & Schreibe, T. (1998). *The University of South Florida word association, rhyme, and word fragment norms*.
<http://w3.usf.edu/FreeAssociation/>
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24 (2), 223–242.
- Nesselhauf, N. (2005). *Collocations in a learner corpus (Studies in Corpus Linguistics, Vol. 14)*. Amsterdam: Benjamins.
- Nippon.com (May 13, 2019). *Year-Round Hiring Aims to Spur More Japanese Students to Learn Overseas*. Retrieved on 10th January 20120 from <https://www.nippon.com/en/japan-data/h00446/year-round-hiring-aims-to-spur-more-japanese-students-to-learn-overseas.html>
- Nouri, N., & Zerhouni, B. (2016). The relationship between vocabulary knowledge and reading comprehension among Moroccan EFL learners. *IOSR Journal of Humanities And Social Science (IOSR-JHSS)* 21 (10) V. 5, 19-26.
Retrieved on 27th January 2020 from:
<https://pdfs.semanticscholar.org/3456/369ab2d82cd6c91dc5ad37def077b91528b2.pdf>
- Nuraeni, C., Carolina, I., Supriyatna, A., Widiati, w., & Bahri, S. (2020). Mobile-Assisted Language Learning (MALL): Students' Perception and Problems towards Mobile.Learning in English Language. *Journal of Physics: Conference Series, 1641* (2020) 012027
doi:10.1088/1742-6596/1641/1/012027
- Obukadeta, P. (2019). *Collocations in a Learner English Corpus: Analysis of Yoruba-speaking Nigerian English Learners' use of Collocations*. [Unpublished doctoral dissertation]. Kingston University, London.
- OECD (2004). *Internationalization and Trade in Higher Education: Opportunities and Challenges*. OECD Publishing Paris.
<https://doi.org/10.1787/9789264015067-en>
- Okada, T. (1999). Samantha Error Corpus.[online] Retrieved on 3 August 2020
<http://www.intcul.tohoku.ac.jp/okada/corpora/Samantha/Samantha-top.html>
- Okada, T. (2002). Remarks on Japanese poor spellers of English. *Yamagata English Studies* 7, 21- 41.
- Okada, T. (2004). A corpus analysis of spelling errors made by Japanese EFL writers *Yamagata English Studies* 9, 17-36.
- Ooi, V., Tan, P., & Chiang, A. (2007). Analyzing personal weblogs in Singapore English: The Wmatrix approach. *Studies in Variation, Contacts and Change in English*, 2.

- Orahood, T., Pearson, D., & Kruze, I. (2004). The impact of student abroad on business students' career goals. *Frontiers 10*, 117-130. Retrieved on January 15, 2020 from: <http://files.eric.ed.gov/fulltext/EJ891452.pdf>
- Ota, H.(2011). Naze kaigai ryuugaku hanare wa okotteirunoka? [Why are Japanese students choosing not to study abroad?] *Kyouiku to Igaku 691*, (1) 68–77.
- Pagán, A., Megan Bird, M., Hsiao, Y., and Nation, K. (2020). Both semantic diversity and frequency influence children's sentence reading. *Scientific Studies of Reading 24* (4), 356-364.
DOI: 10.1080/10888438.2019.1670664
- Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL programme. *TESL Canada 11*, 9-27. <https://doi.org/10.18806/tesl.v11i1.623>
- Parker, G. (1999). Evaluating study abroad programs. *Proceedings of JALT 99 Conference: Teacher belief, teacher action connecting research and the classroom*, 132-135.
- Partington, A., (2012). Corpus analysis of political language. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell, Oxford.
- Pérez-Vidal, C. (2014). Study abroad and formal instruction contrasted: The SALA project. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts*. (pp.17-58). Amsterdam, Netherlands: John Benjamins.
- Pérez-Vidal, C., Juan-Garau, M., Mora, J., & Valls-Ferrer, M. (2012). Oral and written development in formal instruction and study abroad: Differential effects of learning context. In C. Muñoz (Ed.), *Intensive Exposure Experiences in Second Language Learning*, (pp.213-233). Bristol: Multilingual Matters.
- Peters, E. (2016). The learning burden of collocations: The role of interlexical and intra lexical factors. *Language Teaching Research 20* (1), 113-138.
doi: 10.1177/1362168814568131
- Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC® listening, reading, speaking, and writing tests to predicting performance on real-life English language tasks. *Language Testing 32*, 151–167.
- Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review 56* (2), 282-307.
- Qian, D. (2002). Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning 52* (3), 513–536.

- Racine, J., Higginbotham, G., & Munby, I. (2014). Exploring non-native norms: A new direction in word association research. *Verb 3* (2), 13-15.
- Ramanan, K. (2016). *Types of free variation in the writing of Sri Lankan ESL language learners*. Retrieved on 5 August 2020 from Language in India www.languageinindia.com 16:4 April 2016 ISSN 1930-2940
- Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* [Unpublished doctoral dissertation]. Lancaster University, U.K.
- Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004, January). *The UCREL semantic analysis system*. [Conference: Workshop] Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC'04 Lisbon, Portugal. Retrieved on March 15, 2022 from: https://www.researchgate.net/publication/228881331_The_UCREL_semantic_analysis_system#fullTextFileContent
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics 13* (4), 519–549.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge *Language Testing 10* (3), 355–371.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Rees, J., & Klapper, J. (2008). Issues in the quantitative longitudinal measurement of second language progress in the study abroad context. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 89–105). London: Routledge.
- Regan, V. (1998). Sociolinguistics and Language Learning in a Study Abroad Context. *Frontiers: The Interdisciplinary Journal of Study Abroad, 4*, (2), 61-90.
- Revier, R., & Henriksen, B. (2006). Teaching collocations. Pedagogical implications based on a cross-sectional study of Danish EFL. In M. Bendtsen, M. Björklund, C. Fant and L. Forsman (Eds.) *Språk, lärande och utbildning i sikte Pedagogiska fakulteten Åbo Akademi Vasa* (pp.191-206).
- Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly 10* (1), 77-89. doi:10.2307/3585941
- Romaine, S. (2003). Variation. In C.J. Doughty and M.H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 409–435). Oxford: Blackwell.
- Ryan, J., & Lafford, B. (1992). Acquisition of Lexical Meaning in a Natural Environment: "Ser" and "Estar" and the Granada Experience. *Hispania, 75* (3), 714-722.

- Sanz, C., & Morales-Front, A. (2018) Introduction: Issues in study abroad research and practice. In C. Sanz and A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (pp.1-16). Taylor and Francis Group, Abingdon, UK.
- Sasaki, M. (2004). A multiple-data analysis of the 3.5-year development of EFL student writers *Language Learning* 54 (3), 525-582.
- Sasaki, M. (2007). Effects of study-abroad experiences on EFL writers: A multiple-data analysis. *Modern Language Journal* 91 (4), 602-620.
- Sasaki, M. (2009). Changes in English as a foreign language students' writing over 3.5 years: A socio-cognitive account. In R. M. Manchón (Ed.) *Writing in foreign language contexts: Learning, teaching, and research* (pp. 49-76). Clevedon UK: Multilingual Matters.
- Sasaki, M. (2011). Effects of varying lengths of study-abroad experiences on Japanese EFL students' L2 writing ability and motivation: A longitudinal study. *TESOL Quarterly* 45 (1), 81-105.
- Sasaki, M. (2016). L2 writers in study-abroad contexts. In R. Manchón and P. Matsuda (Eds.) *Handbook of second and foreign language writing* (pp. 161-180). Boston: De Gruyter Mouton.
- Savicki, V. (2013). The effects of affect on study abroad students. *Frontiers: The Interdisciplinary Journal of Study Abroad* 22 (1), 131-147.
- Schmitt, N. (1998a). Measuring collocational knowledge: Key issues and an experimental assessment procedure. *I.T.L. Review of Applied Linguistics* 119-120, 27-47.
- Schmitt, N. (1998b). Tracking the incremental acquisition of second language vocabulary: a longitudinal study *Language Learning* 48 (2), 281-317.
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Schmitt, N. (2010) *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan
- 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.
- Scott, G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. (2019). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavioural Research* 51, 1258-1270.
- Scott, M. R. (2010). *WordSmith Tools help manual (Version 5.0)*. Liverpool: Lexical Analysis Software.

- Scott, M.R. and C. Tribble. (2006). *Keywords and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in home and study abroad contexts. *Studies in Second Language Acquisition*, 26, 173-199.
- Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review - Revue Canadienne Des Langues Vivantes - CAN MOD LANG REV* 68, 138-163.
- Shapiro, J. J. (2011). *The Language of Suicide Notes*. [Unpublished doctoral dissertation]. The University of Birmingham, UK.
<http://etheses.bham.ac.uk/1525/>
- Shehata, A. (2008). *L1 influence on the reception and the production of collocations by advanced ESL/EFL Arabic learners of English* [Published doctoral thesis]. Ohio University, USA.
- Simon, E., & Fell, P. (2012). Using mobile learning resources in foreign language instruction. *EDUCAUSE Review*. Retrieved on February 14, 2020 from: <https://er.educause.edu/articles/2012/6/using-mobile-learning-resources-in-foreign-language-instruction>
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J., Jones, S., & Daley, R. (1970). English lexical studies. In R. Krishnamurthy (Ed.), *English collocation studies: The OSTI report* (pp. 2-204). London: Continuum.
- Singleton, D. (1999). *Exploring the Second Language Mental Lexicon*. Cambridge: Cambridge University Press.
- Siyanova, A., & Schmitt, N. (2008). L2 Learner production and processing of collocation: A multi study perspective. *The Canadian Modern Language Review* 64 (3), 429-458.
- Shimmi, Y., & Ota, H. (2018). Super-short-term study abroad in Japan: A dramatic increase. *International Higher Education* 94, 13-15. Retrieved on January 10, 2020 from <https://doi.org/10.6017/ihe.2018.0.10559>
- Shimmi, Y. (2012). The problematic decline of Japanese international students. *International Higher Education* 64, 9-10. Retrieved January 8, 2020 from <https://ejournals.bc.edu/index.php/ihe/article/view/8558/7691>
- Smith, J., & Schmidt, D. (1996). Variability in written Japanese: towards a socio-linguistics of script choice. *Visible Language* 30, 46-71.

- Smith, S. (2017, December 6) *New reports reveals 4.6 million students studying abroad in 2017*. Retrieved 1 January 2020 from: <https://www.studydestinations.com/new-report-reveals-4-6-million-students-studying-abroad-in-2017/>
- Snow, M., Padilla, A., & Campbell, R. (1988). Patterns of Second Language Retention of Graduates of a Spanish Immersion Program. *Applied Linguistics* 9, 182-196.
- Sprenger-Charolles, L., Siegel, L., Jiménez, J., & Ziegler, J. (2011). Prevalence and reliability of phonological, surface and mixed profiles in dyslexia: A review of studies conducted in languages varying in orthographic depth. *Scientific Studies of Reading* 15 (6), 498–521.
- Stewart, J., Batty, A. O., & Bovee, N. (2012). Comparing multidimensional and continuum models of vocabulary acquisition: An empirical examination of the Vocabulary Knowledge Scale. *TESOL Quarterly* 46 (4), 695–721.
- Storch, N., & Hill, K. (2008) What happens to international students' English after one semester at university? *Australian Review of Applied Linguistics* 31 (1), 4.1-4.17.
- Sugiura, R., Imai, N., Hamilton, M., Dean, E., & Ashcroft, R. (2020). Input and output in Japanese high school government-approved English textbooks. *J. Higher Education, Tokai University (Hokkaido Campus)* 21, 1-16.
- Suzuki-Parker, J., & Higginbotham, G. (2019). Does method of administration influence word association test responses? *Vocabulary Education Research Bulletin* 8 (1), 11–16.
- Tanaka, N., & Manning, C. (2018). Examining trends and future directions for short-term study abroad from a stakeholder's perspective. *Shimane University Comprehensive Policy Theory* 35, 1-11.
- TOEFL (2019). *TOEFL iBT tests: test and score data summary for: January 2017-December 2017*. Retrieved January 8, 2020 from https://www.ets.org/s/toefl/pdf/94227_unlweb.pdf
- Towell, R., Hawkins, N., & Bazergui, N. (1996). The Development of Fluency in Advanced Learners of French. *Applied Linguistics*, 17, (1), 84–119.
- Tracy-Ventura, N., Dewaele, J.M., Köylü, Z., & McManus, K. (2016). Personality changes after the 'Year Abroad'?: A mixed-methods study. *Study Abroad Research in Second Language Acquisition and International Education* 1 (1), 107-126.
- Tremblay, D. (1966). Laws of learning general and specialized vocabulary. In *Proceedings of the 74th Annual Convention of the American Psychological Association* 1 (pp. 229–30). Washington DC: APA.

- Tuladhar, A., & Akatsuka, M. (2017). Influence of accurate pronunciation on correctness of spelling in written English: Does accurate pronunciation lead to correct spelling? *Bulletin of the Faculty of Contemporary International Studies, Nagoya University of Foreign Studies* 13, 97-112.
- Van Berkel, A., (2005). The role of the phonological strategy in learning to spell in English as a second Language. In V. Cook. and B. Bassetti (Eds.), *Second Language Writing Systems* (pp 97-121). Clevedon: Multilingual Matters Ltd.
- Van Der Zee, K. I., & Van Oudenhoven, J. P. (2000). The Multicultural Personality Questionnaire: A multidimensional instrument of multicultural effectiveness. *European Journal of Personality*, 14 (4), 291-309.
- Van Leeuwen, M. (2014). Systematic stylistic analysis. In: Kaal, B., Maks, I. and Elfrinkhof, A. (eds) *From Text to Political Positions: Text Analysis across Disciplines*. Amsterdam: John Benjamins, 225–244.
- Ward, C., Bochner, S., & Furnham, A. (2001). *The psychology of culture shock* (2nd ed.). Routledge.
- Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly* 9 (2), 172-185.
- Wang, M. and Geva, E. (2003). Spelling performance of Chinese children using English as a second language: Lexical and visual-orthographic processes. *Applied Psycholinguistics* 24 (1), 1-25.
- Weltens, B., Van Els, T., and Schils, E. (1989). The long-term retention of French by Dutch students. *Studies in Second Language Acquisition* 11 (2), 205-216.
- Wesche, M., and Paribakht, T. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review* 53 13–40.
- West, C. (2015). Japan looks to take flight. *International Educator Japan Supplement* 2 (16).
- Willis, R. (2017). Taming the Climate? Corpus analysis of politicians' speech on climate change. *Environmental Politics* 26 (2), 212-231. DOI: 10.1080/09644016.2016.1274504
- Wister, O., (1907). *How Doth the Simple Spelling Bee*. [e-book] Norwood: The Macmillan Company. Retrieved on 12 September 2020 from Project Gutenberg. <http://www.gutenberg.org/ebooks/23923>
- Wongsarnpigoon, I. (2018). Vocabulary in junior high school textbooks and exams. In P. Clements, A.Krause, & P. Bennett (Eds.), *Language teaching in a global age: Shaping the classroom, shaping the world*. Tokyo: JALT.

- Wolhuter, C. (2016). Jullien: Founding Father of Comparative and International Education Still Pointing the Way. *Education Provision to Every One: Comparing Perspectives from Around the World BCES Conference Books*, 14 (1), 19-24.
Retrieved on 1 January 2022 from <https://files.eric.ed.gov/fulltext/ED568118.pdf>
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press
- Wray, A. (2009). Conclusion: Navigating L2 collocation research. In A. Barfield and H. Gyllstad (Eds.), *Researching Collocations in Another language: Multiple Interpretations* (pp.232-244). Basingstoke: Palgrave Macmillan.
- Yifan, Yu. (2021 November 15). International students to U.S. rebound only 4% after COVID-hit year. *Nikkei Asia*. Retrieved on 1 January 2022 from <https://asia.nikkei.com/Business/Education/International-students-to-U.S.-rebound-only-4-after-COVID-hit-year>
- Yonezawa, A.(2010). Much ado about ranking: Why can't Japanese universities internationalize? *Japan Forum* 22 (1-2), 121–137.
- Zareva, A., Schwanenflugel, P., and Y. Nikolova. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition* 27, 567-595.
- Zareva, A. (2012). Partial word knowledge: Frontier words in the L2 mental lexicon. *International Review of Applied Linguistics in Language Teaching (IRAL)* 50, 277–301.
- Zareva, A. (2014). Frontier words: L1 and L2 partial knowledge of vocabulary. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. DOI: 10.1002/9781405198431.wbeal1434
- Zaytseva, V. (2016). *Vocabulary acquisition in study abroad and formal instruction: an investigation on oral and written lexical development* [Unpublished doctoral dissertation], Universitat Pompeu Fabra, Spain. Retrieved from <http://hdl.handle.net/10803/387120> accessed 1 January, 2022
- Zaytseva, V., Pérez-Vidal, C., and Miralpeix, I. (2018). Vocabulary acquisition during Study Abroad (SA): A comprehensive review of the research. In C. Sanz and A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (pp.210-225). Taylor and Francis Group, Abingdon, UK.

Appendices

Appendix 1: Replication study Lex30 Scoring protocols

This is a list of protocols followed when scoring the Lex30 test. Words from the JACET 1000 list were used. Every answer which occurred on this list was awarded '0' points. Answers which occurred which were not on this list were awarded '1' point. The number of misspelled (but acceptable) words and discounted words were also noted.

Unacceptable

1. No proper nouns to be counted: *Japan, Canada, McDonalds, Kentucky*
2. No numbers
3. No acronyms to be counted: *USA DVD PC CM MC*
4. Japanese words – even those which appear to be an approximation of English words. Eg: *anime* unless they satisfy condition (5) below.
5. Prompt words (used as word association responses): A problem described in Jimenez Catalán and Moreno Espinosa (2005) P.41. However prompt words are acceptable as long as they satisfy condition (3) below.
6. Where two words are written for a single entry only one word will be noted. If one or both words are from L2+ category a maximum of one point will be credited.
For example:
Pot - *hot water* (counted as one word – credited with no points as both words are 1K level)
Potato – *dietary fiber* (counted as one word – credited with one point even though both words are from L2+ category).

Acceptable

1. Misspelled but still recognizable words, although Jiménez Catalán and Moreno Espinosa (2005) argue against this saying that there should be greater score weighting for correctly spelled answers.
2. Each response to the test was lemmatized so that:
 - (i) Responses with an inflectional suffix (plural forms, past tenses, comparatives)
 - (ii) Frequent, regular derivational affixes (-able, -ly)were counted as base-forms of these words. These criteria correspond to levels 2 and 3 of Bauer and Nation's "Word Families" (Bauer & Nation 1993)
3. The same answers written down for different prompt words as long as there is some kind of semantic relationship:
furniture – *bed* and rest - *bed* seat – *movie* television – *movie*
furniture – *sofa* seat – *sofa*

4. Some credit is given when a two part 'phrasal verb' is written instead of a single answer. This occasionally happened when students forgot about test instructions. Phrasal verbs were counted as one word. For example:

Hold – *take place* (scores 1 point) or Habit – *get up* (scores 1 point)

5. Words which are on the 'List of English words of Japanese origin' from Wikipedia.
Example: *sushi* but not *Kanji* or *hiragana*

Appendix 2: Example of Lex30 marking protocol

Name Number

Test Three 22

You need 15 minutes to do this test. The test has 30 items. Date 6th March 2017
 For each item, write four words which you think are related.

Example: animal elephant tiger farm wild

attack	heart	boal	gand	kick
board	study	pen	chook	write
close	window	door	shop	textbook
cloth	pants	socks	jacket	T-shirts
dig
dirty	garbage	bakteria	litter	old things
disease	heart	kidney	cancer	medicine
experience	caria	work	job	oneself
fruit	apple	banana	strawberry	graps
furniture	nutrition	baby	married	family
habit
hold	sport
hope	peace	money	nutrition
kick	boal	soccer	ragby	football
map	papper	difficult	convenience
obey	police	rule
pot	hot	water	tea	green tea
potato	hot	soft	fluffy	chips
real
rest	water	tired	sleep	relax
rice	curry	indica	delicious
science	dangerous	medicine	water	elective
seat	sit	stand up	wheelchair	baby car
spell	one	individual
substance	sugar	salt	pepper
stupid	station	park
television	intersting	elective	magagine	sport
tooth	blush	clear	beautiful	sweet
trade	tea	money	bort	flag
window	close	open	sky	star

Total Count 92 Misspelled (OK) = 4 Discounted = 5 1K words = 44 L2K+ words 43

Appendix 3: SPSS Lex30 Analysis : Total number of words

Descriptive Statistics			
	Mean	Std. Deviation	N
Time 1 Total Words	45.55	17.387	38
Time 2 Total Words	52.16	19.692	38
Time 3 Total Words	71.18	22.646	38
Time 4 Total Words	68.13	20.903	38

Tests of Within-Subjects Effects							
Measure: MEASURE_1							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	17450.493	3	5816.831	66.624	<.001	.643
	Greenhouse-Geisser	17450.493	2.375	7346.175	66.624	<.001	.643
	Huynh-Feldt	17450.493	2.549	6845.299	66.624	<.001	.643
	Lower-bound	17450.493	1.000	17450.493	66.624	<.001	.643
Error(Time)	Sphericity Assumed	9691.257	111	87.309			
	Greenhouse-Geisser	9691.257	87.892	110.264			
	Huynh-Feldt	9691.257	94.323	102.746			
	Lower-bound	9691.257	37.000	261.926			

Pairwise Comparisons						
Measure: MEASURE_1						
(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	-6.605*	2.041	.015	-12.296	-.915
	3	-25.632*	2.573	<.001	-32.805	-18.458
	4	-22.579*	2.590	<.001	-29.799	-15.359
2	1	6.605*	2.041	.015	.915	12.296
	3	-19.026*	1.593	<.001	-23.466	-14.586
	4	-15.974*	2.004	<.001	-21.559	-10.388
3	1	25.632*	2.573	<.001	18.458	32.805
	2	19.026*	1.593	<.001	14.586	23.466
	4	3.053	1.877	.674	-2.179	8.284
4	1	22.579*	2.590	<.001	15.359	29.799
	2	15.974*	2.004	<.001	10.388	21.559
	3	-3.053	1.877	.674	-8.284	2.179

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Appendix 4: SPSS Lex30 Analysis : Lex30 score (1K+ words)

Descriptive Statistics			
	Mean	Std. Deviation	N
Time 1 Raw Lex30 Score	18.76	7.145	38
Time 2 Raw Lex30 Score	20.92	8.270	38
Time 3 Raw Lex30 Score	28.79	9.612	38
Time 4 Raw Lex30 Score	26.55	8.026	38

Tests of Within-Subjects Effects							
Measure: Lex30score							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Sphericity Assumed	2512.651	3	837.550	34.870	<.001	.485
	Greenhouse-Geisser	2512.651	2.700	930.469	34.870	<.001	.485
	Huynh-Feldt	2512.651	2.933	856.555	34.870	<.001	.485
	Lower-bound	2512.651	1.000	2512.651	34.870	<.001	.485
Error(time)	Sphericity Assumed	2666.099	111	24.019			
	Greenhouse-Geisser	2666.099	99.915	26.684			
	Huynh-Feldt	2666.099	108.537	24.564			
	Lower-bound	2666.099	37.000	72.057			

Pairwise Comparisons						
Measure: Lex30score						
(I) time	(J) time	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	-2.158	1.107	.353	-5.244	.928
	3	-10.026 [*]	1.299	<.001	-13.646	-6.406
	4	-7.789 [*]	1.140	<.001	-10.967	-4.612
2	1	2.158	1.107	.353	-.928	5.244
	3	-7.868 [*]	1.016	<.001	-10.702	-5.035
	4	-5.632 [*]	1.167	<.001	-8.885	-2.378
3	1	10.026 [*]	1.299	<.001	6.406	13.646
	2	7.868 [*]	1.016	<.001	5.035	10.702
	4	2.237	.990	.179	-.522	4.995
4	1	7.789 [*]	1.140	<.001	4.612	10.967
	2	5.632 [*]	1.167	<.001	2.378	8.885
	3	-2.237	.990	.179	-4.995	.522

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Appendix 5: SPSS Collocation Analysis : Overall Changes

Descriptive Statistics			
	Mean	Std. Deviation	N
Time1	5.71	2.894	38
Time2	7.24	3.044	38
Time3	9.05	3.654	38
Time4	8.68	3.129	38

Tests of Within-Subjects Effects							
Measure: MEASURE_1							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	264.763	3	88.254	16.782	.000	.312
	Greenhouse-Geisser	264.763	2.725	97.153	16.782	.000	.312
	Huynh-Feldt	264.763	2.963	89.355	16.782	.000	.312
	Lower-bound	264.763	1.000	264.763	16.782	.000	.312
Error(Time)	Sphericity Assumed	583.737	111	5.259			
	Greenhouse-Geisser	583.737	100.833	5.789			
	Huynh-Feldt	583.737	109.633	5.324			
	Lower-bound	583.737	37.000	15.777			

Pairwise Comparisons						
Measure: MEASURE_1						
(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	-1.526 [*]	.468	.014	-2.830	-.223
	3	-3.342 [*]	.604	.000	-5.025	-1.659
	4	-2.974 [*]	.490	.000	-4.338	-1.609
2	1	1.526 [*]	.468	.014	.223	2.830
	3	-1.816 [*]	.549	.013	-3.346	-.286
	4	-1.447	.538	.064	-2.948	.053
3	1	3.342 [*]	.604	.000	1.659	5.025
	2	1.816 [*]	.549	.013	.286	3.346
	4	.368	.497	1.000	-1.016	1.753
4	1	2.974 [*]	.490	.000	1.609	4.338
	2	1.447	.538	.064	-.053	2.948
	3	-.368	.497	1.000	-1.753	1.016

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Appendix 6: SPSS Collocation Analysis : Higher Proficiency Group

Descriptive Statistics			
	Mean	Std. Deviation	N
Time1	7.05	3.064	19
Time2	8.68	3.215	19
Time3	10.21	3.809	19
Time4	9.79	3.425	19

Tests of Within-Subjects Effects							
Measure: MEASURE_1							
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	113.303	3	37.768	6.305	.001	.259
	Greenhouse-Geisser	113.303	2.859	39.634	6.305	.001	.259
	Huynh-Feldt	113.303	3.000	37.768	6.305	.001	.259
	Lower-bound	113.303	1.000	113.303	6.305	.022	.259
Error(Time)	Sphericity Assumed	323.447	54	5.990			
	Greenhouse-Geisser	323.447	51.457	6.286			
	Huynh-Feldt	323.447	54.000	5.990			
	Lower-bound	323.447	18.000	17.969			

Pairwise Comparisons						
Measure: MEASURE_1						
(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	-1.632	.754	.265	-3.865	.602
	3	-3.158*	.873	.012	-5.744	-.572
	4	-2.737*	.816	.021	-5.154	-.319
2	1	1.632	.754	.265	-.602	3.865
	3	-1.526	.743	.328	-3.727	.675
	4	-1.105	.820	1.000	-3.534	1.323
3	1	3.158*	.873	.012	.572	5.744
	2	1.526	.743	.328	-.675	3.727
	4	.421	.750	1.000	-1.802	2.645
4	1	2.737*	.816	.021	.319	5.154
	2	1.105	.820	1.000	-1.323	3.534
	3	-.421	.750	1.000	-2.645	1.802

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Appendix 7: SPSS Collocation Analysis : Lower Proficiency Group

	Mean	Std. Deviation	N
Time1	4.37	2.006	19
Time2	5.79	2.070	19
Time3	7.89	3.178	19
Time4	7.58	2.411	19

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	154.355	3	51.452	10.794	.000	.375
	Greenhouse-Geisser	154.355	2.359	65.438	10.794	.000	.375
	Huynh-Feldt	154.355	2.737	56.386	10.794	.000	.375
	Lower-bound	154.355	1.000	154.355	10.794	.004	.375
Error(Time)	Sphericity Assumed	257.395	54	4.767			
	Greenhouse-Geisser	257.395	42.459	6.062			
	Huynh-Feldt	257.395	49.275	5.224			
	Lower-bound	257.395	18.000	14.300			

(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	-1.421	.574	.141	-3.121	.279
	3	-3.526*	.856	.004	-6.063	-.990
	4	-3.211*	.560	.000	-4.869	-1.552
2	1	1.421	.574	.141	-.279	3.121
	3	-2.105	.823	.119	-4.544	.334
	4	-1.789	.712	.130	-3.898	.319
3	1	3.526*	.856	.004	.990	6.063
	2	2.105	.823	.119	-.334	4.544
	4	.316	.671	1.000	-1.673	2.305
4	1	3.211*	.560	.000	1.552	4.869
	2	1.789	.712	.130	-.319	3.898
	3	-.316	.671	1.000	-2.305	1.673

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Appendix 8: Spelling decline examples

<i>Cue</i>	<i>Attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Attack</i>	[volleyball]	volleyball	volleyball	volleyball	valleyball
<i>Board</i>	[black]	black	black	brack	
<i>Cloth</i>	[coat]		coat	coat	cort
<i>Disease</i>	[stomachache]	stomachache	stomachache	stomach ache	stmachache
	[cough]	cough	cough	cough	cohgu
<i>Fruit</i>	[peach]	peach	peach	peeche	peeche
<i>Furniture</i>	[desk]	desk	desk	desk	dest
	[desk]		desk	dest	besk
<i>Hope</i>	[peace]	peace	peace	peace	pease
<i>Kick</i>	[boxing]	boxing	boxcing	boxcthing	boxthing
<i>Obey</i>	[kingdom]		kingdom	kingdom	kingdam
<i>Pot</i>	[water]		water	water	warter
<i>Potato</i>	[vegetable]	vegetable		vegetable	vegetable
<i>Rice</i>	[ball]	ball	ball	ball	boal

Appendix 8: Spelling decline examples - continued

<i>Cue</i>	<i>Attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Seat</i>	[chair]	chair	chair	chair	cheair
<i>Spell</i>	[vocabulary]	vocabulary	vocabulary	vocabulary	vocabarary
	[pencil]	pencil	pencil		pencill

Appendix 9: Spelling decline then improvement examples

<i>Cue</i>	<i>Attempted</i>	<i>Spelling versions produced</i>			
		1	2	3	4
<i>Attack</i>	[heart]	heart	heart	hart	heart
<i>Board</i>	[school]	school	sholl	school	school
<i>Disease</i>	[headache]	headache	head ache	head ache	headache
	[headache]	headache	headache	headach	headache
<i>Fruit</i>	[apple]	apple	apple	aplee	apple
	[melon]	melon	meron		melon
<i>Map</i>	[picture]	picture	picture		picture
	[paper]		paper	papper.	paper
	[destination]	destination	distination		destination
<i>Obey</i>	[metabolic]	metabolic	metarolic	metabolic	metabolic
<i>Potato</i>	[fry]		fry	friy	fry
<i>Science</i>	[subject]	subject	sunject	subject	subject
<i>Seat</i>	[chair]	chair	chear	chair	chair
<i>Spell</i>	[different]	different	deifferent	different	different

Appendix 9: Spelling decline then improvement examples - continued

<i>Cue</i>	<i>Attempted</i>	<i>Spelling versions produced</i>			
		1	2	3	4
<i>Television</i>	[drama]	drama	drama	dorama	drama
	[interesting]	interesting	interesting	intersting	interesting
	[comedy]		comedy	commedy	comedy
<i>Tooth</i>	[white]		white	whaite	white
	[brash]		brash	brashi	brash
<i>Trade</i>	[import]	import	inport		import
	[import]	import	inport	import	import
<i>Window</i>	[open]	open	open	opn	open

Appendix 10: Spelling improvement then decline examples

<i>Cue</i>	<i>Attempted</i>	<i>Spelling versions produced</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>Board</i>	[black]	brack	black		brack
<i>Close</i>	[door]	doar	door	door	doar
<i>Disease</i>	[stomach]	stomache	stomache	stomach	stomache
<i>Fruit</i>	[orange]	orange	orange	orange	orange
	[pineapple]	pinapple	pinapple	pineapple	pinapple
<i>Furniture</i>	[chair]	chair	cheir	chair	cheir
<i>Pot</i>	[water]	worter		water	worter
<i>Potato</i>	[fried]	frid	fried	fried	freied
	[vegetable]		begitable	vegetable	vagitable
<i>Rice</i>	[delicious]	dericious	delicious	delicious	delicio
<i>Science</i>	[dangerous]	dengerous		dangerous	dengiorous
<i>Television</i>	[drama]	dorama	drama	drama	dorama

Appendix 11 : USAS Semantic Tagset

<http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf>

USAS Semantic Tagset

See <http://ucrel.lancs.ac.uk/usas/> for more details.

A GENERAL & ABSTRACT TERMS	I MONEY & COMMERCE	S1.1.1 General
A1 General	I1 Money generally	S1.1.2 Reciprocity
A1.1.1 General actions, making etc.	I1.1 Money: Affluence	S1.1.3 Participation
A1.1.2 Damaging and destroying	I1.2 Money: Debts	S1.1.4 Deserve etc.
A1.2 Suitability	I1.3 Money: Price	S1.2 Personality traits
A1.3 Caution	I2 Business	S1.2.1 Approachability and Friendliness
A1.4 Chance, luck	I2.1 Business: Generally	S1.2.2 Avarice
A1.5 Use	I2.2 Business: Selling	S1.2.3 Egoism
A1.5.1 Using	I3 Work and employment	S1.2.4 Politeness
A1.5.2 Usefulness	I3.1 Work and employment: Generally	S1.2.5 Toughness; strong/weak
A1.6 Physical/mental	I3.2 Work and employment: Professionalism	S1.2.6 Sensible
A1.7 Constraint	I4 Industry	S2 People
A1.8 Inclusion/Exclusion	K ENTERTAINMENT, SPORTS & GAMES	S2.1 People: Female
A1.9 Avoiding	K1 Entertainment generally	S2.2 People: Male
A2 Affect	K2 Music and related activities	S3 Relationship
A2.1 Affect: Modify, change	K3 Recorded sound etc.	S3.1 Relationship: General
A2.2 Affect: Cause/Connected	K4 Drama, the theatre & show business	S3.2 Relationship: Intimate/sexual
A3 Being	K5 Sports and games generally	S4 Kin
A4 Classification	K5.1 Sports	S5 Groups and affiliation
A4.1 Generally kinds, groups, examples	K5.2 Games	S6 Obligation and necessity
A4.2 Particular/general, detail	K6 Children's games and toys	S7 Power relationship
A5 Evaluation	L LIFE & LIVING THINGS	S7.1 Power, organizing
A5.1 Evaluation: Good/bad	L1 Life and living things	S7.2 Respect
A5.2 Evaluation: True/false	L2 Living creatures generally	S7.3 Competition
A5.3 Evaluation: Accuracy	L3 Plants	S7.4 Permission
A5.4 Evaluation: Authenticity	M MOVEMENT, LOCATION, TRAVEL & TRANSPORT	S8 Helping/hindering
A6 Comparing	M1 Moving, coming and going	S9 Religion and the supernatural
A6.1 Comparing: Similar/different	M2 Putting, taking, pulling, pushing, transporting &c.	T TIME
A6.2 Comparing: Usual/unusual	M3 Movement/transportation: land	T1 Time
A6.3 Comparing: Variety	M4 Movement/transportation: water	T1.1 Time: General
A7 Definite (+ modals)	M5 Movement/transportation: air	T1.1.1 Time: General: Past
A8 Seem	M6 Location and direction	T1.1.2 Time: General: Present; simultaneous
A9 Getting and giving; possession	M7 Places	T1.1.3 Time: General: Future
A10 Open/closed; Hiding/Hidden; Finding; Showing	M8 Remaining/stationary	T1.2 Time: Momentary
A11 Importance	N NUMBERS & MEASUREMENT	T1.3 Time: Period
A11.1 Importance: Important	N1 Numbers	T2 Time: Beginning and ending
A11.2 Importance: Noticeability	N2 Mathematics	T3 Time: Old, new and young; age
A12 Easy/difficult	N3 Measurement	T4 Time: Early/late
A13 Degree	N3.1 Measurement: General	W THE WORLD & OUR ENVIRONMENT
A13.1 Degree: Non-specific	N3.2 Measurement: Size	W1 The universe
A13.2 Degree: Maximizers	N3.3 Measurement: Distance	W2 Light
A13.3 Degree: Boosters	N3.4 Measurement: Volume	W3 Geographical terms
A13.4 Degree: Approximators	N3.5 Measurement: Weight	W4 Weather
A13.5 Degree: Compromisers	N3.6 Measurement: Area	W5 Green issues
A13.6 Degree: Diminishers	N3.7 Measurement: Length & height	X PSYCHOLOGICAL ACTIONS, STATES & PROCESSES
A13.7 Degree: Minimizers	N3.8 Measurement: Speed	X1 General
A14 Exclusion/particularizers	N4 Linear order	X2 Mental actions and processes
A15 Safety/Danger	N5 Quantities	X2.1 Thought, belief
B THE BODY & THE INDIVIDUAL	N5.1 Entirely; maximum	X2.2 Knowledge
B1 Anatomy and physiology	N5.2 Exceeding; waste	X2.3 Learn
B2 Health and disease	N6 Frequency etc.	X2.4 Investigate, examine, test, search
B3 Medicines and medical treatment	O SUBSTANCES, MATERIALS, OBJECTS & EQUIPMENT	X2.5 Understand
B4 Cleaning and personal care	O1 Substances and materials generally	X2.6 Expect
B5 Clothes and personal belongings	O1.1 Substances and materials generally: Solid	X3 Sensory
C ARTS & CRAFTS	O1.2 Substances and materials generally: Liquid	X3.1 Sensory: Taste
C1 Arts and crafts	O1.3 Substances and materials generally: Gas	X3.2 Sensory: Sound
E EMOTIONAL ACTIONS, STATES & PROCESSES	O2 Objects generally	X3.3 Sensory: Touch
E1 General	O3 Electricity and electrical equipment	X3.4 Sensory: Sight
E2 Liking	O4 Physical attributes	X3.5 Sensory: Smell
E3 Calm/Violent/Angry	O4.1 General appearance and physical properties	X4 Mental object
E4 Happy/sad	O4.2 Judgement of appearance (pretty etc.)	X4.1 Mental object: Conceptual object
E4.1 Happy/sad: Happy	O4.3 Colour and colour patterns	X4.2 Mental object: Means, method
E4.2 Happy/sad: Contentment	O4.4 Shape	X5 Attention
E5 Fear/bravery/shock	O4.5 Texture	X5.1 Attention
E6 Worry, concern, confident	O4.6 Temperature	X5.2 Interest/boredom/excited/energetic
F FOOD & FARMING	P EDUCATION	X6 Deciding
F1 Food	P1 Education in general	X7 Wanting; planning; choosing
F2 Drinks	Q LINGUISTIC ACTIONS, STATES & PROCESSES	X8 Trying
F3 Cigarettes and drugs	Q1 Communication	X9 Ability
F4 Farming & Horticulture	Q1.1 Communication in general	X9.1 Ability: Ability, intelligence
G GOVT. & THE PUBLIC DOMAIN	Q1.2 Paper documents and writing	X9.2 Ability: Success and failure
G1 Government, Politics & elections	Q1.3 Telecommunications	Y SCIENCE & TECHNOLOGY
G1.1 Government etc.	Q2 Speech acts	Y1 Science and technology in general
G1.2 Politics	Q2.1 Speech etc: Communicative	Y2 Information technology and computing
G2 Crime, law and order	Q2.2 Speech acts	Z NAMES & GRAMMATICAL WORDS
G2.1 Crime, law and order: Law & order	Q3 Language, speech and grammar	Z0 Unmatched proper noun
G2.2 General ethics	Q4 The Media	Z1 Personal names
G3 Warfare, defence and the army: Weapons	Q4.1 The Media: Books	Z2 Geographical names
H ARCHITECTURE, BUILDINGS, HOUSES & THE HOME	Q4.2 The Media: Newspapers etc.	Z3 Other proper names
H1 Architecture, kinds of houses & buildings	Q4.3 The Media: TV, Radio & Cinema	Z4 Discourse Bin
H2 Parts of buildings	S SOCIAL ACTIONS, STATES & PROCESSES	Z5 Grammatical bin
H3 Areas around or near houses	S1 Social actions, states & processes	Z6 Negative
H4 Residence	S1.1 Social actions, states & processes	Z7 If
H5 Furniture and household fittings		Z8 Pronouns etc.
		Z9 Trash can
		Z99 Unmatched

Appendix 12 : UCREL CLAWS7 Tagset

<https://ucrel.lancs.ac.uk/claws7tags.html>

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that), in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g. any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these, those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number, neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction, neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NUU	unit of measurement, neutral for number (e.g. in, cc)
NUU1	singular unit of measurement (e.g. inch, centimetre)

Appendix 12 : UCREL CLAWS7 Tagset - continued

NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g. about, in)
RPK	prep. adv., catenative (about in be about to)
RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be ..)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)

Appendix 12 : UCREL CLAWS7 Tagset - continued

VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

Appendix 13 : Differences at part-of-speech level between the Taiwanese and British discourse in BATTICC (after Lin 2014, p. 311)

Rank	POS code	Taiwanese participants		British participants		Overuse or underuse	LL	POS
		Freq.	%	Freq.	%			
1	VBDZ	9	0.13	128	1.07	–	67.36	was
2	UH	381	5.58	368	3.07	+	66.02	interjections
3	NN1	755	11.05	950	7.93	+	45.52	sing common nouns
4	VVN	19	0.28	119	0.99	–	35.23	past participle of lexical verb
5	PPH1	100	1.47	333	2.78	–	35.01	neuter personal pronoun: it
6	RR21	4	0.06	62	0.53	–	34.62	general adverbs – ditto tags
7	VVD	38	0.56	142	1.19	–	19.53	past tense of lexical verbs
8	VV0	268	3.93	338	2.82	+	16.05	base form of lexical verb
9	VBDR	2	0.03	29	0.24	–	15.76	were
10	VBZ	207	3.03	265	2.21	+	15.38	is

Appendix 13.2 : Example of most frequent Items in BATTICC-O and CANELC. (after Lin 2017, p. 288).

	BATTICC-O	BATTICC-O		CANELC	CANELC	
		Freq.	Percent		Freq.	Percent
1	<i>I</i>	1,829	5.64	<i>the</i>	20,374	3.96
2	<i>and</i>	982	3.03	<i>to</i>	12,300	2.39
3	<i>is</i>	980	3.02	<i>a</i>	11,647	2.26
4	<i>my</i>	915	2.82	<i>of</i>	9,934	1.93
5	<i>to</i>	853	2.63	<i>and</i>	9,749	1.89
6	<i>the</i>	750	2.31	<i>I</i>	8,071	1.57
7	<i>a</i>	552	1.70	<i>in</i>	7,331	1.42
8	<i>in</i>	516	1.59	<i>it</i>	5,423	1.05
9	<i>you</i>	438	1.35	<i>is</i>	5,272	1.02
10	<i>like</i>	415	1.28	<i>for</i>	5,259	1.02

Appendix 13.3 : The ten most significant differences between BATTICC-O and CANELC at the semantic level (after Lin 2017, p. 294).

Rank	Sem. Code	BATTICC-O		CANELC.		Use	LL	Semantic domain
		Freq.	Percent	Freq.	Percent			
1	Z8	4,676	16.08	1,868	9.10	+	981.85	Pronouns
2	P1	554	1.97	470	0.23	+	979.30	Education in General
3	Z99	889	2.91	1,125	5.48	–	586.83	Unmatched
4	F1	572	1.91	1,025	0.50	+	455.08	Food
5	E2+	467	1.57	787	0.38	+	435.55	Like
6	K1	356	1.23	481	0.23	+	429.57	Entertainment in General
7	I1	25	0.09	1,980	0.96	–	325.64	Money Generally
8	S3.1	225	0.78	258	0.13	+	310.26	Personal Relationships
9	K2	205	0.71	235	0.11	+	283.38	Music and Related Activities
10	A13.3	492	1.70	1,566	0.76	+	197.27	Degree: Boosters

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison

R. Code	Time 2		Time 3		+/- use	LL	Semantic Group
	Freq	%.	Freq	%			
1. L1-	4	0.20	0	0.00	-	6.88	Dead
2. T1.1.2	2	0.10	14	0.52	+	6.79	Time: Present; simultaneous
3. P1	23	1.17	58	2.16	+	6.72	Education in general
4. X7+	6	0.31	24	0.90	+	6.70	Wanted
5. S2	7	0.36	24	0.90	+	5.33	People
6. B2	3	0.15	0	0.00	-	5.16	Health and disease
7. A8	0	0.00	4	0.15	+	4.40	Seem
8. Z99	11	0.56	30	1.12	+	4.24	Unmatched
9. A2.1+	2	0.10	10	0.37	+	3.63	Change
10. X2.3+	2	0.10	0	0.00	-	3.44	Learning
11. S7.4+	2	0.10	0	0.00	-	3.44	Allowed
12. N3.6	2	0.10	0	0.00	-	3.44	Measurement: Area
13. I1.3+	2	0.10	0	0.00	-	3.44	Expensive
14. X2.4	0	0.00	3	0.11	+	3.30	Investigate, examine, test
15. N5	0	0.00	3	0.11	+	3.30	Quantities
16. I1.1	0	0.00	3	0.11	+	3.30	Money and pay
17. M7	10	0.51	26	0.97	+	3.27	Places
18. L3	59	3.00	58	2.16	-	3.12	Plants
19. X4.2	4	0.20	1	0.04	-	2.98	Mental object: Means, method
20.T1.1.1	1	0.05	6	0.22	+	2.58	Time: Past
21.A5.1+	11	0.56	7	0.26	-	2.57	Evaluation: Good
22.L2	32	1.63	29	1.08	-	2.54	Live creatures: animals, birds
23.A13.3	5	0.25	2	0.07	-	2.43	Degree: Boosters
24. S9	2	0.10	8	0.30	+	2.23	Religion and the supernatural
25. A5.1-	2	0.10	8	0.30	+	2.23	Evaluation: Bad
26.X5.2-	0	0.00	2	0.07	+	2.20	Uninterested/unenergetic
27.W2-	0	0.00	2	0.07	+	2.20	Darkness
28. S3.1	0	0.00	2	0.07	+	2.20	Personal relationship: General
29. S1.2.4-	0	0.00	2	0.07	+	2.20	Impolite
30. N5.2+	0	0.00	2	0.07	+	2.20	Exceed; waste
31. N3.7+	0	0.00	2	0.07	+	2.20	Long, tall and wide
32. N3.3+	0	0.00	2	0.07	+	2.20	Distance: Far
33. I1.3	0	0.00	2	0.07	+	2.20	Money: Cost and price
34. A2.2	0	0.00	2	0.07	+	2.20	Cause&Effect/Connection
35. A1.5.1	0	0.00	2	0.07	+	2.20	Using
36. B1	57	2.90	99	3.69	+	2.16	Anatomy and physiology

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison - continued

37. Q2.1	3	0.15	10	0.37	+	2.12	Speech: Communicative
38. Q1.3	9	0.46	6	0.22	-	1.89	Telecommunications
39. N5.1-	3	0.15	1	0.04	-	1.76	Part
40. A10-	3	0.15	1	0.04	-	1.76	Closed; Hiding/Hidden
41. X9.2+	1	0.05	0	0.00	-	1.72	Success
42. T2++	1	0.05	0	0.00	-	1.72	Time: Beginning
43. T1.3+	1	0.05	0	0.00	-	1.72	Time period: long
44. S1.1.4+	1	0.05	0	0.00	-	1.72	Deserving
45. O4.6	1	0.05	0	0.00	-	1.72	Temperature
46. N3.5+	1	0.05	0	0.00	-	1.72	Weight: Heavy
47. I1.2	1	0.05	0	0.00	-	1.72	Money: Debts
48. A6.2-	1	0.05	0	0.00	-	1.72	Comparing: Unusual
49. A1.3+	1	0.05	0	0.00	-	1.72	Cautious
50. Q4.3	20	1.02	18	0.67	-	1.64	Media: TV, Radio, Cinema
51. E3+	9	0.46	20	0.75	+	1.56	Calm
52. T3+	4	0.20	2	0.07	-	1.44	Time: Old; grown-up
53. S1.2.5+	4	0.20	2	0.07	-	1.44	Tough/strong
54. F1	210	10.68	256	9.55	-	1.44	Food
55. O4.3	81	4.12	93	3.47	-	1.27	Colour and colour patterns
56. N3.2-	6	0.31	4	0.15	-	1.26	Size: Small
57. M2	9	0.46	7	0.26	-	1.25	Putting, pulling, pushing
58. E4.1+	26	1.32	26	0.97	-	1.25	Happy
59. O1	8	0.41	6	0.22	-	1.24	Substances and materials
60. K5.1	49	2.49	54	2.01	-	1.16	Sports
61. W3	24	1.22	24	0.90	-	1.15	Geographical terms
62. B2-	19	0.97	35	1.31	+	1.14	Disease
63. I3.1	4	0.20	10	0.37	+	1.13	Work and employment
64. Y1	8	0.41	17	0.63	+	1.12	Science and technology
65. X7-	1	0.05	4	0.15	+	1.12	Unwanted
66. X3.5	1	0.05	4	0.15	+	1.12	Sensory: Smell
67. T3--	1	0.05	4	0.15	+	1.12	Time: New and young
68. S1.1.1	1	0.05	4	0.15	+	1.12	Soc actions, states, processes
69. E3-	11	0.56	22	0.82	+	1.12	Violent/Angry
70. A1.1.2	1	0.05	4	0.15	+	1.12	Damaging and destroying
71. Z8	0	0.00	1	0.04	+	1.10	Pronouns
72. Z6	0	0.00	1	0.04	+	1.10	Negative
73. Z4	0	0.00	1	0.04	+	1.10	Discourse Bin
74. Z3	0	0.00	1	0.04	+	1.10	Other proper names
75. X9.1+	0	0.00	1	0.04	+	1.10	Able/intelligent

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison - continued

76. X3.2+	0	0.00	1	0.04	+	1.10	Sound: Loud
77. X2.5+	0	0.00	1	0.04	+	1.10	Understanding
78. X2.2-	0	0.00	1	0.04	+	1.10	No knowledge
79. W5	0	0.00	1	0.04	+	1.10	Green issues
80. T2+	0	0.00	1	0.04	+	1.10	Time: Beginning
81. T1.3-	0	0.00	1	0.04	+	1.10	Time period: short
82. S7.1--	0	0.00	1	0.04	+	1.10	No power
83. S1.2.5-	0	0.00	1	0.04	+	1.10	Weak
84. S1.2.4+	0	0.00	1	0.04	+	1.10	Polite
85. S1.1.3-	0	0.00	1	0.04	+	1.10	Non-participating
86. S1.1.3+	0	0.00	1	0.04	+	1.10	Participating
87. Q4	0	0.00	1	0.04	+	1.10	The Media
88. N3.8-	0	0.00	1	0.04	+	1.10	Speed: Slow
89. N3.8+	0	0.00	1	0.04	+	1.10	Speed: Fast
90. L1	0	0.00	1	0.04	+	1.10	Life and living things
91. I1.3-	0	0.00	1	0.04	+	1.10	Cheap
92. I1.1-	0	0.00	1	0.04	+	1.10	Money: Lack
93. I1.1+	0	0.00	1	0.04	+	1.10	Money: Affluence
94. G2.2-	0	0.00	1	0.04	+	1.10	Unethical
95. F3	0	0.00	1	0.04	+	1.10	Smoking / non-medical drugs
96. F1-	0	0.00	1	0.04	+	1.10	Lack of food
97. E2-	0	0.00	1	0.04	+	1.10	Dislike
98. E2++	0	0.00	1	0.04	+	1.10	Like
99. E1	0	0.00	1	0.04	+	1.10	Emotional Actions, States
100. A4.2-	0	0.00	1	0.04	+	1.10	General
101. A1.9	0	0.00	1	0.04	+	1.10	Avoiding
102. A1.8+	0	0.00	1	0.04	+	1.10	Inclusion
103. A1.7+	0	0.00	1	0.04	+	1.10	Constraint
104. O4.2-	3	0.15	8	0.30	+	1.07	Appearance: Negative
105. A4.1	2	0.10	6	0.22	+	1.04	Kinds, groups, examples
106. Q1.2	40	2.03	44	1.64	-	0.96	Paper documents and writing
107. T1.3	12	0.61	11	0.41	-	0.90	Time: Period
108. X3.1	11	0.56	10	0.37	-	0.86	Sensory: Taste
109. O1.2	27	1.37	46	1.72	+	0.86	Substances / materials: Liquid
110. G1.1	11	0.56	10	0.37	-	0.86	Government
111. T1	8	0.41	16	0.60	+	0.81	Time
112. H2	55	2.80	64	2.39	-	0.74	Parts of buildings
113. T3-	2	0.10	1	0.04	-	0.72	Time: New and young
114. E6-	2	0.10	1	0.04	-	0.72	Worry

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison - continued

115. X2.2+	14	0.71	14	0.52	-	0.67	Knowledgeable
116. S8-	6	0.31	5	0.19	-	0.66	Hindering
117. B5	89	4.53	108	4.03	-	0.66	Clothes / personal belongings
118. W2	3	0.15	7	0.26	+	0.64	Light
119. F2	28	1.42	31	1.16	-	0.64	Drinks and alcohol
120. S5+	3	0.15	2	0.07	-	0.63	Belonging to a group
121. O3	3	0.15	2	0.07	-	0.63	Electricity, electric equipment
122. M4	3	0.15	2	0.07	-	0.63	Sailing, swimming, etc.
123. A5.3+	3	0.15	2	0.07	-	0.63	Evaluation: Accurate
124. L1+	4	0.20	3	0.11	-	0.62	Alive
125. O4.6+	43	2.19	50	1.86	-	0.58	Temperature: Hot / on fire
126. S8+	2	0.10	5	0.19	+	0.57	Helping
127. S2.2	2	0.10	5	0.19	+	0.57	People: Male
128. N2	2	0.10	5	0.19	+	0.57	Mathematics
129. X3.3	1	0.05	3	0.11	+	0.52	Sensory: Touch
130. X2.6+	1	0.05	3	0.11	+	0.52	Expected
131. N5.1+	1	0.05	3	0.11	+	0.52	Entire; maximum
132. A9	1	0.05	3	0.11	+	0.52	Getting ,giving; possession
133. Q2.2	5	0.25	10	0.37	+	0.51	Speech acts
134. B4	24	1.22	27	1.01	-	0.47	Cleaning and personal care
135. M3	53	2.70	64	2.39	-	0.43	Vehicles and transport on land
136. Y2	4	0.20	8	0.30	+	0.41	Information tech / computing
137. X5.2+	9	0.46	16	0.60	+	0.41	Interested/excited/energetic
138. M1	22	1.12	25	0.93	-	0.39	Moving, coming and going
139. X3.1+	17	0.86	19	0.71	-	0.35	Tasty
140. K2	8	0.41	14	0.52	+	0.32	Music and related activities
141. W4	12	0.61	20	0.75	+	0.31	Weather
142. O4.2+	16	0.81	26	0.97	+	0.31	Judge appearance: Positive
143. B3	10	0.51	17	0.63	+	0.31	Medicines, medical treatment
144. N4	3	0.15	6	0.22	+	0.30	Linear order
145. A5.2+	3	0.15	6	0.22	+	0.30	Evaluation: True
146. N5+	6	0.31	6	0.22	-	0.29	Quantities: many/much
147. N3.3-	15	0.76	17	0.63	-	0.27	Distance: Near
148. K5.2	19	0.97	22	0.82	-	0.27	Games
149. A12-	15	0.76	17	0.63	-	0.27	Difficult
150. O2	55	2.80	82	3.06	+	0.26	Objects generally
151. O4.6-	5	0.25	9	0.34	+	0.25	Temperature: Cold
152. C1	5	0.25	9	0.34	+	0.25	Arts and crafts
153. S3.2	5	0.25	5	0.19	-	0.24	Relations: Intimacy and sex

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison - continued

154. S1.2.3+	5	0.25	5	0.19	-	0.24	Selfish
155. O4.5	7	0.36	12	0.45	+	0.24	Texture
156. I2.2	26	1.32	40	1.49	+	0.23	Business: Selling
157. X9.2-	2	0.10	4	0.15	+	0.20	Failure
158. K4	17	0.86	20	0.75	-	0.20	Drama, theatre, show business
159. I1	9	0.46	10	0.37	-	0.20	Money generally
160. H4	9	0.46	10	0.37	-	0.20	Residence
161. A5.2-	2	0.10	4	0.15	+	0.20	Evaluation: False
162. X2	4	0.20	4	0.15	-	0.19	Mental actions and processes
163. S2.1	4	0.20	4	0.15	-	0.19	People: Female
164. M5	4	0.20	4	0.15	-	0.19	Flying and aircraft
165. H1	13	0.66	15	0.56	-	0.19	Architect, houses, buildings
166. A7+	4	0.20	4	0.15	-	0.19	Likely
167. M8	12	0.61	19	0.71	+	0.17	Stationary
168. O4.4	4	0.20	7	0.26	+	0.16	Shape
169. E4.1-	4	0.20	7	0.26	+	0.16	Sad
170. G3	8	0.41	13	0.48	+	0.15	War, defence, army, weapons
171. E2+	6	0.31	10	0.37	+	0.15	Like
172. A1.4	3	0.15	3	0.11	-	0.14	Chance, luck
173. Q4.1	11	0.56	13	0.48	-	0.12	The Media: Books
174. S4	11	0.56	17	0.63	+	0.11	Kin
175. X3.2	2	0.10	2	0.07	-	0.10	Sensory: Sound
176. W1	27	1.37	34	1.27	-	0.10	The universe
177. S7.4-	1	0.05	2	0.07	+	0.10	Not allowed
178. O1.2-	1	0.05	2	0.07	+	0.10	Dry
179. O1.1	24	1.22	30	1.12	-	0.10	Substances / materials: Solid
180. N3.5	1	0.05	2	0.07	+	0.10	Measurement: Weight
181. N1	2	0.10	2	0.07	-	0.10	Numbers
182. M6	9	0.46	14	0.52	+	0.10	Location and direction
183. F4	1	0.05	2	0.07	+	0.10	Farming & Horticulture
184. A9-	1	0.05	2	0.07	+	0.10	Giving
185. A6.3+	2	0.10	2	0.07	-	0.10	Comparing: Varied
186. A5.4-	1	0.05	2	0.07	+	0.10	Evaluation: Unauthentic
187. A12+	2	0.10	2	0.07	-	0.10	Easy
188. A1.5.2+	2	0.10	2	0.07	-	0.10	Useful
189. A1.2+	1	0.05	2	0.07	+	0.10	Suitable
190. X3.4	3	0.15	5	0.19	+	0.08	Sensory: Sight
191. S7.1+	5	0.25	8	0.30	+	0.08	In power
192. A6.1-	3	0.15	5	0.19	+	0.08	Comparing: Different

Appendix 14 : Semantic domain keyness - SA T2 corpus v. SA T3 corpora comparison - continued

193. A3+	3	0.15	5	0.19	+	0.08	Existing
194. X4.1	19	0.97	28	1.04	+	0.07	Mental object: Conceptual
195. X1	1	0.05	1	0.04	-	0.05	Psychological Actions, States
196. S6+	1	0.05	1	0.04	-	0.05	Strong obligation or necessity
187. S1.2.1-	1	0.05	1	0.04	-	0.05	Formal/Unfriendly
198. N5-	1	0.05	1	0.04	-	0.05	Quantities: little
199. N3.2+++	1	0.05	1	0.04	-	0.05	Size: Big
200. G2.2+	1	0.05	1	0.04	-	0.05	Ethical
201. G2.1-	1	0.05	1	0.04	-	0.05	Crime
202. E5-	1	0.05	1	0.04	-	0.05	Fear/shock
203. A6.2+	1	0.05	1	0.04	-	0.05	Comparing: Usual
204. A4.2+	1	0.05	1	0.04	-	0.05	Detailed
205. A1.4+	1	0.05	1	0.04	-	0.05	Lucky
206. A9+	8	0.41	12	0.45	+	0.04	Getting and possession
207. A5.3-	5	0.25	6	0.22	-	0.04	Evaluation: Inaccurate
208. Q4.2	17	0.86	22	0.82	-	0.03	The Media: Newspapers etc.
209. H5	59	3.00	78	2.91	-	0.03	Furniture, household fittings
210. A1.1.1	11	0.56	16	0.60	+	0.03	General actions / making
211. X2.1	4	0.20	6	0.22	+	0.02	Thought, belief
212. O1.3	4	0.20	6	0.22	+	0.02	Substances and materials: Gas
213. A10+	27	1.37	38	1.42	+	0.02	Open; Finding; Showing
214. Z5	7	0.36	10	0.37	+	0.01	Grammatical bin
215. X8+	2	0.10	3	0.11	+	0.01	Trying hard
216. Q3	15	0.76	21	0.78	+	0.01	Language, speech, grammar
217. Q1.1	2	0.10	3	0.11	+	0.01	Linguistics, Communication
218. I2.1	2	0.10	3	0.11	+	0.01	Business: Generally
219. Z2	3	0.15	4	0.15	-	0.00	Geographical names
220. T1.1.3	5	0.25	7	0.26	+	0.00	Time: Future
221. O4.1	3	0.15	4	0.15	-	0.00	Appearance, physical props
222. N3.2+	8	0.41	11	0.41	+	0.00	Size: Big
223. K1	11	0.56	15	0.56	-	0.00	Entertainment generally
224. G2.1	3	0.15	4	0.15	-	0.00	Law and order
225. A15-	6	0.31	8	0.30	-	0.00	Danger
226. A11.1+	12	0.61	16	0.60	-	0.00	Important
