Vocabulary Knowledge Components:
Knowledge, acquisition and conceptualisation



Beatriz González-Fernández

M.A., B.A., B.Ed.

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

December 2017

Abstract

Vocabulary knowledge is a multifaceted construct for which complete command comprises mastering various types of knowledge. Achieving this full mastery is deemed challenging and problematic for second language learners, and yet it is key for successful and appropriate language use (Webb & Nation, 2017). By understanding how these various kinds of knowledge behave and relate, practitioners can better systematise when and how they are introduced to the students, and thus facilitate the overall vocabulary learning process. Nevertheless, little is known about how these multiple types of word knowledge are acquired by learners of a second/foreign language. This thesis attempts to contribute to bridging this gap by empirically exploring the knowledge, acquisition and conceptualisation of overall vocabulary knowledge in second languages. Five studies are devoted to this aim.

Study 1 examined the knowledge and order of acquisition of four different components of vocabulary knowledge (form–meaning link, collocations, derivatives and multiple meanings) in recall and recognition. It shows that those components vary considerably in their difficulty for second language (L2) learners, and implicational scaling analysis revealed that a consistent order of acquisition of these components can be established.

Study 2 investigated how various factors typically identified as influencing vocabulary learning affect the mastery of those four word knowledge components. The results suggest that cognateness status of the target words has the strongest effect on knowledge of most vocabulary aspects by Spanish learners of L2 English, followed by frequency and proficiency. Importantly, this cognate influence raised the question of whether the order of acquisition of word knowledge components retrieved in Study 1 was unique to L2 English learners of cognate languages.

Study 3 explored the theoretical conceptualisation of vocabulary knowledge by attempting to provide empirical support for its multidimensionality. Structural equation modelling analyses exposed that the various types of vocabulary knowledge were so interconnected that they could not be considered separate dimensions. This finding suggests that the different aspects of vocabulary knowledge are better seen as various word knowledge difficulty levels that comprise a unidimensional construct.

Finally, Studies 4 and 5 tried to provide generalisability to the previous findings by replicating Studies 1 and 3, respectively, with Chinese L2 learners of English. Study 4 investigated the hypothesis that the order of acquisition outlined in Study 1 for Spanish learners depended partially on the cognate status of the participants' language. The implicational scaling analysis with the non-cognate Chinese population, however, corroborated the order of acquisition of vocabulary aspects previously suggested, indicating that a universal vocabulary acquisition order might exist.

Study 5 also confirmed the finding that vocabulary knowledge is not empirically multidimensional. In addition, it provided evidence for the fact that the vocabulary knowledge construct functions similarly across the Chinese and Spanish learners of English as a unidimensional concept.

Taken together, the results presented in this thesis support the notion that mastering the various components involved in word knowledge is a slow and incremental process, and demonstrate that some of them seem to be consistently acquired before others. Nonetheless, the findings also suggest that these different types of word knowledge cannot be considered separate dimensions in the learners' mental lexicon, and thereby the multidimensional structure of vocabulary knowledge might need to be reconsidered. Overall, while more research is needed, these findings offer useful insights for the learning and teaching of vocabulary in a more systematic and efficient manner.

Acknowledgements

During this PhD journey I have been fortunate to meet and be accompanied by truly remarkable people who have made this experience unforgettable. The support of my friends, teachers and colleagues helped me develop both personally and professionally. I would like to thank them all, because without them this thesis would have not been possible.

First of all, I would like to express my deepest gratitude for my mentor and supervisor, Prof. Norbert Schmitt. It was he who sparked my interest in vocabulary and motivated me to pursue a PhD in the field. He has provided me with invaluable guidance and help throughout the process, and has transmitted to me his enthusiasm for research. I feel privileged to be his last PhD student, and in the name of all his former supervisees, I would like to thank him for his encouragement and all we learnt from him.

I am also grateful to my second supervisor, Prof. Zoltán Dörnyei, for his assistance and believing in me before, during, and after the PhD.

I would also like to thank all members of the Vocabulary Research Group during my time at Nottingham: Hana Almutairi, Samuel Barclay, Dr. Mélodie Garnier, Benjamin Kremmel, Duyen Le Thi, Dr. Marijana Macis, Dr. Ana Pellicer-Sánchez, Niloofar Rahimi, Dr. Michael P.H. Rodgers, Dr. Kholood Saigh, Dr. Pawel Szudarski, and Dr. Laura Vilkaitė. The constructive group discussions, their ambition and their critical minds have inspired me to become a better researcher. My thanks go also to the members of the Psycholinguistics Reading Group, particularly Dr. Kathy Conklin, Dr. Walter van Heuven and Dr. Dominic Thompson, for their invaluable insights about psycholinguistics.

I also want to thank Dr. Ali H. Al-Hoorie and Dr. Fabio Parente for their patient help and invaluable comments about statistics.

My sincere appreciation also goes to all my participants in Spain and China whose tireless impetus made this work possible. Also to the teachers who allowed me to engage with their students, including Dr. Laura Llanos and Dr. Margaret Dowens. In particular, I am grateful to Prof. María Teresa Llamazares Prieto, for being the first person that saw the research potential in me, her unfailing moral support and for encouraging me to make the most of myself as a professional (and for the *croquetas*!).

I too want to thank Prof. Stuart Webb for hosting me during my institutional visit in Canada and for his critical feedback on my research and practical words of advice and inspiration.

Very special thanks go to my beloved friends in Nottingham, Laura and Ben, with whom I not only share the interest for vocabulary, but also a lasting friendship. They always were there for me whenever I needed them, both in person and virtually. Their invaluable support throughout this process and our trips will never be forgotten.

Thanks also go to my office partners Awatif Alruwaili, Aya Van Renterghem and Xiaofan Xu, and to my MA friend Dr. Gavin Brookes, for the therapeutic discussions and advice, especially during the final stages of this thesis.

My very exceptional gratitude goes to Matthew Fielding for his restless patience and support during the most challenging times of my PhD. His special nature makes me feel fortunate, and challenges me to be a better person every day.

Finally, I am grateful to my family in Spain: my mother Patro, my father Tomás, my sister Alicia, my brother-in-law Pablo, and my nephew Alejo and niece Pepa, who always get my best smile. They have at all times believed in me and supported me, even without always understanding. This thesis is dedicated to them. *A ellos va dedicada esta tesis. ¡Gracias!*

Declaration

I declare that the work presented in this thesis is my own and was conducted during my time as a PhD student at the University of Nottingham. Parts of this thesis have been published or submitted to peer-reviewed journals:

- A fragment of the literature review outlined in Chapter 2 was published in co-authorship with Norbert Schmitt in *The Routledge Handbook of Instructed Second Language Acquisition* (González-Fernández & Schmitt, 2017)
- Studies 1 and 3, presented in Chapters 4 and 6, have been combined and submitted as a paper co-authored with Norbert Schmitt, and is currently under review
- The studies presented in Chapters 5, 7 and 8 are currently being prepared for submission

The chapters in this thesis are more detailed versions of these published papers. The published versions should be used for any citations or page reference.

Table of Contents

Abstract i
Acknowledgements iii
Declaration v
Table of Contents vi
List of Figures xii
List of Tables xiv
List of Abbreviations xviii

Chapter 1: Introduction 1
1.1. Structure of the thesis 2

Chapter 2: Vocabulary Knowledge: The construct under review 6
2.1. Conceptualising Vocabulary Knowledge 6
2.1.1. What it means to know vocabulary 6
2.1.2. Theoretical approaches to vocabulary knowledge 13
2.1.3. Summary of the conceptualisation issues 26
2.2. Vocabulary Knowledge and SLA 27
2.2.1. Nation's word knowledge components: what do we know so far? 27
2.2.2. How do word knowledge components relate to each other? 45
2.2.3. Multicomponent vocabulary research 47
2.2.4. Factors influencing vocabulary acquisition 51
2.2.5. Measuring vocabulary knowledge 55
2.2.6. Summary of the issues with L2 vocabulary knowledge 59

Chapter 3: Development of the measurement instruments for vocabulary knowledge components 61
3.1. Design considerations 61
3.1.1. Target population 62
3.1.2. Selection of target word knowledge aspects 63
3.1.3. Selection of target words 65
3.1.4. Criteria for stimulus words 67

3.2. Measures 69
3.2.1. Form recall knowledge of the form–meaning link 71
3.2.2. Meaning recognition knowledge of the form–meaning link 72
3.2.3. Form recall knowledge of derivatives 73
3.2.4. Form recognition knowledge of derivatives 74
3.2.5. Meaning recall knowledge of multiple meanings 76
3.2.6. Meaning recognition knowledge of multiple meanings 77
3.2.7. Form recall knowledge of collocations 78
3.2.8. Form recognition knowledge of collocations 79
3.2.9. 'I don't know' option 80
3.2.10. Two scoring perspectives 81
3.2.11. VLT and language background 82
3.3. Piloting 83
3.3.1. Items piloting 83
3.3.2. Order of tests administration 85
3.3.3. Test duration 86
3.4. Cross-test influence 88
3.5. Conclusion 90

Chapter 4: Knowledge and order of acquisition of vocabulary knowledge components by EFL learners: An Implicational Scale 91
4.1. Background of the study 91
4.1.1. What is Implicational Scaling? 92
4.1.2. Implicational Scaling analysis in linguistics 95
4.1.3. Implicational Scaling in SLA research 96
4.2. Purpose and research questions 98
4.3. Methodology 99
4.3.1. Guttman Scaling procedure of goodness-of-fit 99
4.3.2. Cut-off criteria for Implicational Scaling 102
4.3.3. Participants 103
4.3.4. Test-battery administration 105
4.3.5. Data analysis 106
4.4. Results and interpretation 107
4.4.1. General results 107

4.4.2. Implicational scaling results 115
4.5. Discussion 126
4.5.1. Knowledge of word knowledge components 127
4.5.2. The interrelatedness of word knowledge components 128
4.5.3. Order of acquisition of word knowledge components 130
4.5.4. Limitations and directions for further research 137
4.6. Conclusion 139

Chapter 5: Factors influencing knowledge of vocabulary components: Frequency, cognateness and proficiency 140
5.1. Background of the study 140
5.1.1. What makes a word difficult? Factors affecting vocabulary acquisition 140
5.2. Purpose and research questions 148
5.3. Methodology 149
5.3.1. Mixed-effects modelling analysis 150
5.4. Results and interpretation 153
5.4.1. Preliminary analysis: Descriptive statistics and correlations 153
5.4.2. Model analysis: Mixed-effects modelling 155
5.5. Discussion 175
5.5.1. Effect of target word cognateness 175
5.5.2. Effect of target word frequency 180
5.5.3. Effect of proficiency 182
5.5.4. Effect of word length 184
5.5.5. Pedagogical implications 185
5.5.6. Limitations and directions for further research 187
5.6. Conclusion 188

Chapter 6: The vocabulary knowledge construct and the relationships among its components: A Structural Equation Model 190
6.1. Background of the study 190
6.1.1. Multidimensional interpretation of vocabulary knowledge 190
6.1.2. The use of SEM in the examination of vocabulary knowledge 191
6.2. Structural equation modelling methodology 201

6.2.1. Introduction to SEM 201
6.2.2. Basic concepts of SEM 203
6.2.3. Steps in SEM 207
6.2.4. Model evaluation 209
6.3. Purpose and research questions 217
6.4. Methodology 218
6.4.1. The hypothesised model 218
6.4.2. Data analysis 220
6.5. Results and interpretation 222
6.5.1. Descriptive statistics and correlations 222
6.5.2. Structural equation modelling analyses 223
6.6. Discussion 241
6.6.1. Representation of word knowledge aspects 242
6.6.2. Unsupported multidimensionality of vocabulary knowledge 243
6.6.3. Unidimensional conceptualisation of vocabulary knowledge 246
6.6.4. Contribution of components to overall vocabulary knowledge 247
6.6.5. Consideration of terminology 249
6.6.6. Limitations and directions for future research 251
6.7. Conclusion 252

Chapter 7: Order of acquisition of vocabulary knowledge components in L2 English: A replication study 253
7.1. Background of the study 253
7.2. Purpose and research questions 254
7.3. Methodology 255
7.3.1. Participants 255
7.3.2. Chinese translation 257
7.3.3. Piloting 258
7.3.4. Test-battery administration 258
7.3.5. Data analysis 259
7.4. Results and interpretation 259
7.4.1. General results 260
7.4.2. Implicational scaling results 268
7.5. Comparison of results between Spanish and Chinese learners 278

7.5.1. Reliability of tests....278
7.5.2. Correlations among word knowledge components and vocabulary size……. 279
7.5.3. Correlations between the individual words' scores and their raw frequency....281
7.5.4. Knowledge of vocabulary components across groups 282
7.5.5. Implicational scaling results....286
7.6. Discussion....289
7.6.1. Knowledge of word knowledge components 290
7.6.2. Interrelationships between word knowledge components 292
7.6.3. General order of acquisition of word knowledge components 293
7.6.4. Pedagogical implications....296
7.6.5. Limitations and directions for further research 300
7.7. Conclusion 302

Chapter 8: Conceptualising vocabulary knowledge: A replication study 304
8.1. Background of the study....304
8.2. Purpose and research questions 305
8.3. Methodology....306
8.3.1. The hypothesised model....306
8.3.2. Data analysis 308
8.4. Results and interpretation 309
8.4.1. Descriptive statistics and correlations 309
8.4.2. Structural equation modelling analyses....310
8.5. Comparison of results between the Spanish and Chinese groups 327
8.5.1. Measurement instruments with individual words 327
8.5.2. Latent variables for the hypothesised model....328
8.5.3. Hypothesised model 329
8.5.4. Revised model 331
8.5.5. Multigroup CFA of the revised model 334
8.6. Discussion....343
8.6.1. Behaviour of word knowledge aspects and the vocabulary knowledge construct in L2s....343
8.6.2. Equivalence of the vocabulary knowledge construct in L2 learners....345

8.6.3. Implications for understanding and measuring vocabulary knowledge…. 348
8.6.4. Pedagogical implications 350
8.6.5. Limitations and directions for future research 353
8.7. Conclusion 355

Chapter 9: General discussion and conclusion 356
9.1. Main findings and implications 356
9.1.1. Order of acquisition of vocabulary components 356
9.1.2. Factors influencing vocabulary acquisition 359
9.1.3. Conceptualisation of vocabulary knowledge 362
9.2. Directions for future research 364
9.3. Concluding remarks 366

References 368
Appendices 414

List of Figures

Figure 1 A network model of vocabulary (Meara & Wolter, 2004, p. 89). 22

Figure 2 The lexical space (Daller et al., 2007, p. 8). 23

Figure 3 Search parameters for collocate retrieval. 79

Figure 4 Example of implicational matrix. 94

Figure 5 Order retrieved in the implicational scale for the Spanish group with the 80% accuracy level (from easier to more difficult). 116

Figure 6 Order retrieved in the implicational scale for the Spanish group with the 75% accuracy level (from easier to more difficult). 120

Figure 7 Implicational scale of word knowledge components. 130

Figure 8 Interaction between cognateness and proficiency in form–meaning link recall. 159

Figure 9 Interaction between cognateness and proficiency in form–meaning link recognition. 162

Figure 10 Interaction between cognateness and proficiency in derivatives recognition. 166

Figure 11 Interaction between cognateness and proficiency in multiple-meanings recall. 168

Figure 12 Interaction between cognateness and proficiency in multiple-meanings recognition. 170

Figure 13 Example of a structural equation model. 203

Figure 14 Hypothesised model of relationships between word knowledge components 219

Figure 15 Example of CFA for each word knowledge component. 224

Figure 16 Pooled-CFA of all factors in hypothesised model (Spanish group). 228

Figure 17 Model 1: Hierarchical model of Vocabulary Knowledge (Spanish group). 231

Figure 18 Model 2: Recognition and recall knowledge as independent aspects of vocabulary knowledge (Spanish group). 235

Figure 19 Direct effect of form–meaning link on depth of knowledge aspects (Spanish group). 239

Figure 20 Direct effect of recognition on recall masteries (Spanish group). 240

Figure 21 Order retrieved in the implicational scale for the Chinese group with the 80% accuracy level (from easier to more difficult). 269

Figure 22 Order retrieved in the implicational scale for the Chinese group with the 75% accuracy level (from easier to more difficult). 272

Figure 23 Comparison of VLT sections' scores by Chinese and Spanish learners. . 283

Figure 24 Comparison of scores for each aspect of Chinese and Spanish learners. . 285

Figure 25 General order of difficulty of vocabulary components by Chinese and Spanish learners of English (80% accuracy)..... 287

Figure 26 Generalisable implicational scale of word knowledge components. 294

Figure 27 Hypothesised model of relationships between word knowledge components. 307

Figure 28 Pooled-CFA of all factors in hypothesised model (Chinese group). 314

Figure 29 Model 1: Hierarchical model of Vocabulary Knowledge (Chinese group). 317

Figure 30 Model 2: Recognition and recall knowledge as independent aspects of vocabulary knowledge (Chinese group). 321

Figure 31 Direct effect of form–meaning link on depth of knowledge aspects (Chinese group). 325

Figure 32 Direct effect of recognition on recall masteries (Chinese group). 326

Figure 33 Baseline model for multigroup analysis. 336

List of Tables

Table 1 Nation's (2013, p. 49) framework of the components involved in knowing a word. 16

Table 2 Test formats per word knowledge aspect. 70

Table 3 List of target words (ordered by frequency). 84

Table 4 Potential test-administration orders. 85

Table 5 Target words in each test section. 87

Table 6 Proportion of participants mastering a later aspect but not a prior one (80% accuracy). 89

Table 7 Proportion of participants mastering a later aspect but not a prior one (75% accuracy). 90

Table 8 Spanish participants' exposure to English (per week). 104

Table 9 Internal consistency reliability for word knowledge aspects' tests (Spanish group). 108

Table 10 Descriptive statistics of the Spanish participants' performance (raw total scores) 110

Table 11 Descriptive statistics of the Spanish participants' performance (percentage total scores). 111

Table 12 Correlations between word knowledge components (Spanish participants). 114

Table 13 Results of Guttman Procedure with 80% accuracy criterion (Spanish group). 117

Table 14 Mokken scale for 80% accuracy criterion, reliability and dimensionality assessment (Spanish group). 118

Table 15 Results of Guttman Procedure with 75% accuracy criterion (Spanish group). 120

Table 16 Mokken scale for 75% accuracy criterion, reliability and dimensionality assessment (Spanish group). 121

Table 17 Number of participants mastering each aspect (Spanish group). 122

Table 18 Results of Guttman Procedure for individual words (80% accuracy) (Spanish group). 124

Table 19 Range, mean and standard deviation of variables (Lemma Frequency before log-transformation) 151

Table 20 Correlation between words' lemma frequency (COCA) and knowledge in the different aspects (Spanish group). 153

Table 21 Mean percentage score for cognateness status in all aspects. 154

Table 22 Core model of form–meaning recall knowledge. 155
Table 23 Full model of form–meaning recall knowledge. 156
Table 24 Model of form–meaning recall knowledge. 157
Table 25 Main effects on form–meaning recall knowledge. 158
Table 26 Comparison of model fit for the form–meaning recall aspect. 159
Table 27 Comparison of model fit for the form–meaning recognition aspect. 160
Table 28 Model of form–meaning recognition knowledge. 161
Table 29 Comparison of model fit for the derivative recall aspect. 163
Table 30 Model of derivative recall knowledge. 163
Table 31 Comparison of model fit for the derivative recognition aspect. 164
Table 32 Model of derivative recognition knowledge. 165
Table 33 Comparison of model fit for the multiple-meanings recall aspect. 166
Table 34 Model of multiple-meanings recall knowledge. 167
Table 35 Comparison of model fit for the multiple-meanings recognition aspect. 169
Table 36 Model of multiple-meanings recognition knowledge. 170
Table 37 Comparison of model fit for the collocation recall aspect. 171
Table 38 Model of collocation recall knowledge. 171
Table 39 Comparison of model fit for the collocation recognition aspect. 172
Table 40 Model of collocation recognition knowledge. 173
Table 41 Summary of the mixed-effects modelling results. 174
Table 42 Summary of fit indices and their threshold levels. 213
Table 43 Summary of fit indices combinations. 216
Table 44 Descriptive statistics (%) and intercorrelations between components (Spanish group). 223
Table 45 Fit indices and reliability for the CFA models with 20 target words as indicators (Spanish group). 225
Table 46 Pooled-CFA fit indices (Spanish group). 228
Table 47 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for Pooled-CFA model (Spanish group). 229
Table 48 *Z* scores and significance for Pooled-CFA (Spanish group). 230
Table 49 Model 1: Fit indices (Spanish group). 232
Table 50 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for hypothesised model (Spanish group). 232

Table 51 Relative contribution of the different aspects to vocabulary knowledge, *Z* scores and significance (Spanish group). 233

Table 52 Model 2: Fit indices (Spanish group). 235

Table 53 Model comparison for parsimony (Spanish group). 236

Table 54 *Z* scores and significance of word knowledge aspects as indicators of vocabulary knowledge (Model 2) (Spanish group). 237

Table 55 Model fit for Alternative Model 1 (Figure 19) (Spanish group). 240

Table 56 Model fit for Alternative Model 2 (Figure 20) (Spanish group). 241

Table 57 Chinese participants' exposure to English (per week). 257

Table 58 Internal consistency reliability for word knowledge aspects' tests (Chinese group). 260

Table 59 Descriptive statistics of the Chinese participants' performance (raw total scores). 262

Table 60 Descriptive statistics of the Chinese participants' performance (percentage total scores). 263

Table 61 Correlations between word knowledge components (Chinese group). 266

Table 62 Correlation between words' lemma frequency (COCA) and knowledge of the different aspects (Chinese group). 267

Table 63 Results of Guttman Procedure with 80% accuracy criterion (Chinese group). 270

Table 64 Mokken scale for 80% accuracy criterion, reliability and dimensionality assessment (Chinese group). 271

Table 65 Results of Guttman Procedure with 75% accuracy criterion (Chinese group). 273

Table 66 Mokken scale for 75% accuracy criterion, reliability and dimensionality assessment (Chinese group). 274

Table 67 Number of participants mastering each aspect (Chinese group). 274

Table 68 Results of Guttman Procedure for individual words (80% accuracy) (Chinese group). 277

Table 69 Summary of correlation coefficients between measures, means, and standard deviations as a function of language group. 280

Table 70 Number of participants mastering each aspect (whole sample). 288

Table 71 Descriptive statistics (%) and intercorrelations between components (Chinese group). 310

Table 72 Fit indices and reliability for the CFA models with 20 target words as indicators (Chinese group). 311

Table 73 Pooled-CFA fit indices (Chinese group). 314

Table 74 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for Pooled-CFA model (Chinese group). 315

Table 75 *Z* scores and significance for Pooled-CFA (Chinese group). 316

Table 76 Model 1: Fit indices (Chinese group). 318

Table 77 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for hypothesised model (Chinese group). 318

Table 78 Relative contribution of the different aspects to vocabulary knowledge, *Z* scores and significance (Chinese group). 319

Table 79 Model 2: Fit indices (Chinese group). 321

Table 80 Model comparison for parsimony (Chinese group). 322

Table 81 *Z* scores and significance of word knowledge aspects as indicators of vocabulary knowledge (Model 2) (Chinese group). 323

Table 82 Model fit for Alternative Model 1 (Figure 31) (Chinese group). 325

Table 83 Model fit for Alternative Model 2 (Figure 32) (Chinese group). 326

Table 84 Summary of goodness-of-fit statistics for hypothesised model in both groups. 330

Table 85 Summary of goodness-of-fit statistics for revised model in both groups. . 331

Table 86 Summary of goodness-of-fit statistics for test of multigroup measurement invariance. 337

Table 87 Latent mean difference between Spanish and Chinese learners. 342

List of Abbreviations

AIC	Akaike information criterion
AVE	average variance extracted
BIC	Bayes information criterion
BNC	British National Corpus
CFA	confirmatory factor analysis
CFI	comparative fit index
COCA	Corpus of Contemporary American English
CR	composite reliability
EFL	English as a foreign language
ICC	intra-class correlation coefficient
IS	implicational scale
L1	first language
L2	second language
MI	mutual information
ML	maximum likelihood
MLR	maximum likelihood robust
NS	native speaker
PPVT	Peabody Picture Vocabulary Test
RMSEA	root mean square error of approximation
SEM	structural equation modelling
SLA	second language acquisition
SRMR	standardised root mean square residual
VKS	Vocabulary Knowledge Scale
VLT	Vocabulary Levels Test
VST	Vocabulary Size Test
WAT	Word Association Task
WLS	weighted least squares
WLSMV	weighted least squares means and variance adjusted

Chapter 1

Introduction

Vocabulary is an essential element when learning and using a language. Although it was once a neglected aspect of language learning and teaching (Meara, 1983), its relevance is now widely acknowledged not only by teachers and researchers, but also by learners of second and foreign languages (Nation, 1990, 2013; Schmitt, 2010; Webb & Nation, 2017). Vocabulary knowledge is central to all the receptive and productive language skills (i.e., reading, listening, writing, and speaking), and is a main predictor of general language proficiency, both in first and second languages (Alderson, 2007; Laufer & Goldstein, 2004; Milton, 2013). For these reasons, there has been an increase in vocabulary research in the last few decades, and it is now one of the most extensively researched areas in second language teaching, learning, processing and assessment (Dóczi & Kormos, 2016; Nation, 2013). Despite this large amount of research, the field still lacks a generally accepted theory of how vocabulary is acquired.

Researchers point to the great complexity of the vocabulary knowledge construct as the main reason for this lack of an overall acquisition theory (Milton & Fitzpatrick, 2014; Schmitt, 2014). On the one hand, vocabulary knowledge has been interpreted and described in various ways. Therefore, a plethora of conceptualisations exist, which has made this complex construct even more difficult to understand and study. On the other hand, vocabulary knowledge comprises various word knowledge components within it. Even though most descriptions of vocabulary address this multidimensionality, there is no agreement on which specific components it contains. Sometimes different components overlap under the same term, and sometimes the same term is employed to refer to different notions. As a result, even research that measures the same component can arrive at different results if they adopt a different approach to vocabulary knowledge. These issues mean that typically the multifaceted nature of vocabulary has not been attended to by researchers. Most of the attention in research

and the classroom has been placed on the development of size of vocabulary knowledge (i.e., the number of words known). This abundant research on how many words are learnt has led to a clear imbalance in vocabulary studies with respect to how well words are known (i.e., depth of vocabulary knowledge). That is, overall vocabulary knowledge is still largely unexplored. While examining size of vocabulary knowledge is important, only by adopting a more comprehensive approach that takes into account size and depth can we gain a deeper understanding of how overall vocabulary is known and acquired. This more nuanced knowledge is important to help systematise the instruction of vocabulary and to further develop research in the field of vocabulary acquisition.

To this end, the multidimensionality of vocabulary needs to be carefully investigated. It has been argued that a way of starting exploring this construct is by working with the individual aspects that constitute vocabulary knowledge (Fitzpatrick & Milton, 2014; Schmitt, 1997). However, the majority of the research following this approach has generally focused on only one type of word knowledge. Thus, most of our current limited understanding about this multifaceted construct derives from studies that assessed the components individually. As a consequence, it is still unclear how the different word knowledge components are acquired and fit together. In order to arrive at a general explanatory theory of vocabulary acquisition, the multiple components comprised in vocabulary knowledge need to be examined in a detailed and concurrent manner. This thesis seeks to address this research gap in vocabulary knowledge and acquisition studies by simultaneously examining the knowledge that second language learners of English have of various word knowledge aspects and exploring the nature of their relationships in an attempt to model the structure of the vocabulary knowledge construct. The following chapters will investigate issues such as the nature and understanding of vocabulary, as well as word knowledge aspects, their measurement and relationships, and the factors that affect their acquisition.

1.1. Structure of the thesis

This thesis reports on five self-standing but interrelated studies, looking at the knowledge, acquisition and conceptualisation of vocabulary. It begins with an extensive literature review in Chapter 2, aimed at providing a general theoretical background for the scope of the five studies. Chapter 2 starts by presenting and discussing the different

conceptualisations of vocabulary knowledge, and stating the one that will be followed in this thesis. After summarising the research findings on the knowledge and acquisition of individual components comprised in vocabulary knowledge, I argue for the significance and necessity of treating vocabulary knowledge as a multifaceted construct in research by analysing and discussing the multicomponent vocabulary studies. I then provide an overview of the factors that have been found to influence vocabulary acquisition. Finally, the chapter discusses the effects that different descriptions and measures of vocabulary components have on our understanding of the vocabulary knowledge construct.

Chapter 2 provides a general overview of the topics in the thesis, but does not extensively discuss the methodological issues in previous research. This will be addressed in the individual studies, where each of the study chapters have their own background section presenting the relevant research methodologies and focusing exclusively on the research pertaining to the specific study. These are: implicational scaling in Chapter 4, research on factors influencing vocabulary learning in Chapter 5, and structural equation modelling in Chapter 6. (Chapters 7 and 8 replicate the studies in Chapters 4 and 6 respectively, and thus do not repeat the background information).

Chapter 3 outlines the development of the test battery designed to measure knowledge of vocabulary components. Design considerations such as the selection of word knowledge aspects, the target words, and the target population to be assessed are discussed first. Then, the different test formats are presented, along with their scoring protocols. The chapter ends with some piloting analyses and a discussion of the cross-test influence of the test battery.

After Chapter 3, the five study chapters are presented. Each chapter is written up as a standalone study that involves their own research questions, methodology, analyses, results and interpretation of the findings. Yet, since the different studies are interrelated, some connections are made with findings of preceding studies. The chapters start with an introduction that links the different studies together, and then a background section focusing on the issues specific to the particular studies. Subsequently, a detailed description of the methodology employed to answer the research questions is provided. After presenting and analysing the results, they are interpreted in a discussion section. Finally, the pedagogical applications and implications of the results are presented, and the limitations of the study are assessed in the form of suggestions for future research.

Chapter 4 reports on a study which looked at the knowledge and acquisition of four word knowledge components in recall and recognition masteries: form–meaning link, collocations, derivatives and multiple meanings. Using implicational scaling analysis, it shows a systematic order of acquisition of these word knowledge components by Spanish second language learners of English. An argument is presented for replicating the results in order to explore the validity of that order.

Chapter 5 presents a study that investigates the inter- and intralexical factors that affect knowledge of the different word knowledge components tested by Spanish English as a Foreign Language (EFL)[1] learners. Mixed-effect modelling analysis shows the strong influence of cognateness on knowledge of most vocabulary components. The chapter argues for the need to investigate EFL learners of non-cognate languages in order to find out how the order of acquisition of components could have been influenced by this cognateness status.

Chapter 6 explores the multidimensionality of vocabulary knowledge empirically by means of structural equation modelling analyses. It provides an empirically-based conceptualisation of vocabulary knowledge that presents it as a single construct. As in the previous chapter, an argument is made for the need to explore the generalisability of the results by examining other populations of second language (L2) learners.

Chapter 7 reports on a replication of the study presented in Chapter 4 by examining the knowledge that Chinese learners of English had of the same word knowledge components (i.e., form–meaning link, collocations, derivatives and multiple meanings). This chapter presents the results of the Chinese participants first, and then moves into providing a comprehensive comparison of the results in the two populations (i.e., Spanish and Chinese learners). The study in Chapter 7 originated as a need to better understand the findings of Chapters 4 and 5. It builds on and extends the findings to a different population of learners. For this reason, the studies in Chapters 4 and 7 are presented as two different studies.

Chapter 8 replicates the findings of Chapter 6 for Chinese learners of L2 English. After providing the results for this learner population individually, the chapter compares the findings with those for the Spanish EFL learners in Chapter 6. Having found the same underlying conceptualisation of vocabulary knowledge for both groups,

[1] In the present thesis, the terms ‘second language’ (L2) and ‘foreign language’ (FL) are used interchangeably.

the study also argues for the theoretical and empirical equivalence of the vocabulary knowledge construct in EFL learners by employing multigroup analysis of variance.

Finally, Chapter 9 summarises the findings of the previous chapters and discusses general implications for the wider context of vocabulary knowledge. This chapter ends by pointing out directions for further research that is needed to finally arrive at a generally accepted theory of vocabulary acquisition.

Chapter 2

Vocabulary Knowledge: The construct under review

The present chapter provides a general review of the research and frameworks of vocabulary knowledge. As this thesis is devoted to the examination of the conceptualisation and acquisition of overall vocabulary knowledge, this chapter explores those issues in two main sections. The first part looks at what is involved in knowing vocabulary and the different ways vocabulary knowledge has been understood and conceptualised in research. The second section presents an overview of empirical studies that account for the knowledge and acquisition of vocabulary in a second language (L2), the factors that have been found to affect its mastery, and the different ways vocabulary knowledge has been operationalised in research.

2.1. Conceptualising Vocabulary Knowledge

2.1.1. What it means to know vocabulary

Traditionally, learning vocabulary in a L2 involved acquiring some kind of an equivalent translation of single words in the first language (L1). This understanding of vocabulary confined its teaching to the memorisation of bilingual word lists that had to be combined with some type of syntactic rules. However, this notion of vocabulary soon began to be recognised as rather vague and problematic (Palmer, 1917, 1921). For example, following this idea, a Spanish student learning English as a L2 would learn that the English word *do* means 'hacer' in Spanish, and thus, applying syntactical rules, produce the sentence *'I do my bed every morning'[2]. The sentence is grammatical and

[2] The asterisk (*) indicates an incorrect sentence.

the broad meaning of the word is correct; yet, the sentence is inappropriate due to incorrect lexical selection. As this view of word knowledge did not capture the whole nature of the construct, vocabulary began to be understood as a multifaceted concept.

Today, vocabulary knowledge is generally deemed a complex construct that involves more than learning L2 translations of single words. Its very complexity has made it an elusive concept, and as such, it is yet not clear what it means to know vocabulary (Milton & Fitzpatrick, 2014). In an attempt to shed some light into and characterise such a difficult construct, many researchers have proposed different classifications of lexical knowledge. Most of these vocabulary knowledge descriptions are based on the distinction between two generally accepted sets of dimensions of vocabulary knowledge: breadth vs. depth, and receptive vs. productive mastery.

2.1.1.1. Breadth and depth

The first scholars in conceptualising vocabulary knowledge as divided into depth and breadth were Anderson and Freebody (1981). They defined vocabulary breadth as "the number of words for which the person knows at least some of the significant aspects of meaning" and vocabulary depth as "the quality or depth of understanding" (p. 93). Although these notions have been given different terminologies in the literature (see Henriksen, 1999; Meara, 1996; Read, 1993; Wesche & Paribakht, 1996), most researchers nowadays accept *size* or *breadth* as terms to indicate the quantity of words a person knows to a certain extent, and *depth* to refer to how well those words are known.

The notion of vocabulary size was described more precisely by Anderson and Freebody than that of depth. This additional clarity has made vocabulary size the subject of most research on vocabulary knowledge, and this extensive research, in turn, has yielded some clear descriptions of what size of vocabulary knowledge involves. As a result, size is now consensually accepted and generally operationalised as knowledge of the form–meaning link of multiple words (i.e., mapping a given L2 form to its meaning and/or an existing meaning to the appropriate L2 form), although it sometimes has also comprised knowledge of just the spoken and written form of words (Milton, 2009). Interest in investigating vocabulary size has led to the development of numerous tests that tap into this concept. For example, two standardised and widely employed

tests that provide an estimate of vocabulary size are the Vocabulary Size Test (VST) (Nation & Beglar, 2007), a multiple-choice test divided by frequency bands where learners are required to recognise the meaning of a given word form, and the Peabody Picture Vocabulary Test (PPVT) (Dunn & Dunn, 2007), conceptually similar but presenting pictures for learners to select the one that represents the word they hear. Apart from these two tests that were specifically designed to measure the learners' vocabulary size, researchers argue that the majority of the traditional 'vocabulary' tests currently available (e.g., L1-L2 translation tests, recognition of word meaning, or recall of word form) also tap into vocabulary size, since they typically measure the form–meaning link knowledge. Therefore, the concept of vocabulary size has been broadly examined, both directly and indirectly.

This main focus on size of knowledge within vocabulary research means that the study of depth has often been disregarded (Henriksen, 1999; Nation & Webb, 2011; Schmitt, 2010). It has been suggested that the reason for this imbalance on the L2 vocabulary studies in favour of size is due to the intrinsic difficulty in describing and measuring depth, already appreciated in the vague original definition of the construct provided by Anderson and Freebody (1981). While the concept has experienced an increased interest among researchers in recent years, the complexity and ambiguity of defining the degree to which a word is known has led to multiple conceptualisations, and no true single description exists to date (see Section 2.1.2 in this chapter). For example, Read (1993) defines depth as "the quality of the learner's vocabulary knowledge" (p. 357), which materialises in knowledge of synonyms and collocations, while Li and Kirby (2015) define it as the "richness of word knowledge" (p. 1), including, at least, knowledge of the multiple meaning senses of a word, collocations and morphological knowledge. Depth of knowledge can also be understood as the learner's knowledge of links among individual words, and more specifically knowledge of multiple semantic lexical associations with other words in the lexicon (Haastrup & Henriksen, 2000; Meara, 1996). It can also refer to having some degree of proficiency of as many word knowledge dimensions as possible (Nation, 2013). Overall, the multifaceted nature of depth of vocabulary knowledge has made its conceptualisation and operationalisation difficult, and as a result, to date there is not one comprehensive definition accepted by all researchers of what depth of knowledge involves (Milton, 2009).

Despite this lack of consensus on what exactly constitutes vocabulary depth, researchers agree that it is not an all-or-nothing construct, but rather involves a deeper and richer understanding of words and their connections in the mental lexicon (Anderson & Freebody, 1981; Beck, McKeown, & Kucan, 2013; Henriksen, 1999; Qian & Schedl, 2004). Researchers and practitioners also agree upon the importance for learners to develop depth of vocabulary knowledge. While having a large vocabulary size is central to function in English, vocabulary needs to be known at a deeper level of quality than the form–meaning link for learners to use it appropriately, accurately and fluently in varied contexts.

The relationship between size and depth has also been of much research interest. It is generally agreed that the development of depth of word knowledge is more problematic for learners and thus lags behind vocabulary size, regardless of the learners' proficiency level (Laufer, Elder, Hill, & Congdon, 2004; Schmitt, 2014; Wesche & Paribakht, 1996). For example, Webb (2007a) found that size (as represented by the form–meaning link) was generally learned earlier than depth (e.g., syntagmatic associations, paradigmatic associations, and word class). A possible reason is that size tests often measure more explicit knowledge (i.e., form–meaning link), while depth tests measure more implicit aspects, such as associations in the mental lexicon (Schmitt, 2014). This difference in knowledge means that size and depth are often treated as independent dimensions of vocabulary knowledge. Nevertheless, it is not possible to draw a totally clear line between size and depth, since the knowledge required to demonstrate size of vocabulary inevitably involves some quality of knowledge, and vice versa. This intuition is also supported by different studies showing medium to large correlations between size and depth. For example, Li and Kirby (2015), employing hierarchical regressions, found that although a measure of breadth and multiple measures assessing depth of vocabulary loaded into two different factors, the correlation between them was moderate (r = .51). Similarly, a moderate correlation (r = .55) was also found by Tseng and Schmitt (2008), and by Zhang and Yang (2016) (r = .62). These studies suggest that breadth and depth are indeed interrelated and contribute to one another, although they can still be considered two different dimensions.

However, other studies have found much higher correlations between these two concepts. For example, Qian (1999, 2002) found a correlation of between .82 and .70 in L2 learners, Vermeer (2001) a correlation of .85 for L1 learners and .76 for L2

learners, Tannenbaum, Torgesen, & Wagner (2006) a correlation of .88, Gyllstad (2009) correlations of between .83 and .88, or Nguyen and Webb (2017) of around .70. These strong correlations found between size and depth have led some authors to question whether they are indeed distinct vocabulary dimensions, and thus view them as the same construct (Gyllstad, 2013; Nguyen & Webb, 2017; Vermeer, 2001). Indeed, Kieffer and Lesaux (2012a), employing confirmatory factor analysis techniques, found that the breadth and depth dimensions could not be considered discriminant, and thus the differences between them do not reflect measurably distinct dimensions. In this sense, size and depth would be considered to develop in parallel, so that as learners' vocabulary size increases, so does their vocabulary depth. Therefore, the relationship between size and depth is still uncertain, although the aforementioned results seem to indicate that instead of being considered two independent constructs they are better seen as two parts of the same dimension that overlap, interrelate and facilitate each other.

To sum up, the consequence of research having consistently focused on size of vocabulary is a lack of a clear understanding of vocabulary depth. Although it is accepted that word learning is incremental and possibly starts with a vague knowledge of the form–meaning connection, word learning must continue to become more precise and deeper in knowledge.

2.1.1.2. Receptive and productive knowledge

The idea of receptive and productive knowledge (a.k.a passive vs. active mastery) was first conveyed by Palmer (1917, 1921). He proposed that the distinction between them refers to what the learners can do with words in a language. Receptive knowledge denotes the learner being able to understand words encountered while reading or listening, and productive knowledge refers to using words in speaking or writing recalling it from memory without any other support. Hence, these concepts are skilled-based constructs representing the ability to recognise and understand input or produce output (Schmitt, 2014).

The concepts of receptive and productive knowledge have been conceptualised in different ways. For example, Meara (1997) proposes a connection approach, in which the shift from receptive to productive knowledge depends on the number and quality of the connections and networks between lexical items. Henriksen (1999) suggests a

dimensional approach, suggesting that breadth and depth of knowledge are connected to receptive vocabulary knowledge, while the capacity to fluently and naturally access this knowledge for communicative purposes refers to productive knowledge. In her view, receptive and productive mastery would be another dimension of vocabulary knowledge, independent from size and depth but interrelated. Alternatively, Melka (1997) and Read (2000) bring up a different perspective to this relationship, in which productive and receptive knowledge is not seen as a dichotomy, but a continuum. In this approach, learners are believed to first acquire receptive knowledge of a word, and only when sufficient receptive knowledge is acquired, productive knowledge develops, so the receptive knowledge is included and extended in the productive one.

Research on the relationship between these concepts has focused on whether receptive or productive knowledge is acquired before the other, mainly regarding vocabulary size (Aitchison, 2003; Laufer, 1998; Laufer & Paribakht, 1998; Melka, 1997). It has been found that receptive and productive knowledge are learnt incrementally, with receptive mastery being typically reached before productive mastery (Henriksen, 1999; Nation, 2013). This is partly because productive mastery requires knowledge of more word knowledge aspects. Schmitt (2014) suggests that knowing the form–meaning link of a word might be enough for a receptive understanding of that word (although, of course, the more lexical aspects known, the better the comprehension is likely to be). In this situation, the user only needs to recall the meaning attached to the form that has been perceived, since all the other word knowledge aspects (e.g., word class, collocations, grammatical functions) are provided in the context. However, in order to produce a lexical item accurately and appropriately in a specific context, the user needs to know all (or as many as possible) of these aspects. That is, following Read's (2000) approach, productive knowledge is more advanced than receptive knowledge, and thus some receptive lexical knowledge is essential before a word can be used productively.

Different studies testing receptive and productive mastery of just the form–meaning link found that receptive knowledge was higher than productive knowledge (sometimes even five times higher: Nemati, 2010). For example, Vermeer (2001) found that for L2 Dutch learners the scores in a form–meaning recognition task was almost four times as high as in the recall task. This pattern of productive mastery lagging behind receptive mastery was also confirmed by Laufer and Paribakht (1998) for Israeli and

Canadian EFL learners, and by Webb (2008) for Japanese EFL learners, and they found it was especially true for low-frequency words, where learners could only recognise the form. These studies have also found that receptive and productive knowledge of the form–meaning link correlates strongly (e.g., r = .75 (Vermeer, 2001) or r = .72-89 (Laufer & Paribakht, 1998)), providing support for the claim that they are better seen as a continuum than a dichotomy.

There are few studies measuring receptive and productive gains in various word knowledge aspects. Schmitt and Meara (1997) studied the receptive and productive knowledge of two components, and found that participants had higher receptive than productive knowledge in the both aspects tested, although no stable pattern was found in the growths. In an innovative study, Webb (2007a), using non-words, examined the acquisition of receptive and productive mastery of five word knowledge aspects. He found that participants experienced gains in both receptive and productive knowledge, although the receptive knowledge of all the word knowledge aspects was always larger than the productive knowledge. This study was replicated by Chen and Truscott (2010) using real words, and the results were consistent. Therefore, the receptive knowledge of the different word knowledge aspects seems to be more robust and be acquired earlier than productive knowledge. This has led to the belief that speakers of a language will only develop productive knowledge of a limited number of words known receptively (Nation, 2013).

Although it is broadly accepted that there is a distinction between receptive and productive mastery (Laufer & Goldstein, 2004; Webb, 2008) the difference in knowledge between these two dimensions is still not clear. Some studies claim that the gap between receptive and productive knowledge is really big, and therefore they are totally different constructs. For example, Laufer (2005) found that only between 16% and 35% of the vocabulary tested was known both productively and receptively at the 5,000 and 2,000 frequency levels, respectively. The results are more optimistic in other research. Meara (1997) reviewed several studies which suggest that the difference between receptive and productive knowledge is really small, with as much as 92% of the receptive vocabulary being known productively. In 2008, Webb found a similar pattern, with participants having partial productive knowledge of most of the words they knew receptively (77% with strict scoring and 93% with sensitive scoring). This disagreement regarding how much receptive vocabulary is also known productively is

due to the different ways of conceptualising receptive and productive knowledge, emphasising the fact that describing and measuring the concepts of receptive and productive knowledge of a language is confusing and difficult. Our understanding regarding the relationship between those dimensions is also limited by the fact that researchers have typically focused on assessing receptive knowledge of words. Even when both types of knowledge have been researched, most studies have concentrated on the receptive and productive vocabulary size, disregarding the aspects of depth of word knowledge. Thus, more examination on these dimensions and how they relate to each other for the different types of vocabulary knowledge is needed for a clear and adequate definition of the constructs.

2.1.2. Theoretical approaches to vocabulary knowledge

Relying on the concepts of breadth/depth and receptive/productive knowledge, vocabulary research has emphasised the existence of different degrees and types of knowledge involved in vocabulary, which have led to the development of multiple conceptual classifications to understanding the nature of vocabulary knowledge. These can be categorised in three main approaches: *components approach*, *developmental approach*, and *lexical networks approach*. In this section, I will start by discussing the most widely followed approach to vocabulary knowledge in research (i.e., components approach), which is also the one employed in this thesis, to then briefly present other alternatives.

2.1.2.1. Components approach

This approach (a.k.a. *dimensions approach*) divides vocabulary knowledge into its component parts, establishing comparisons between them. It originated from the idea that, in order to better understand the broad vocabulary knowledge construct, it needs to be broken down into smaller, more manageable units. The principal representative of this approach nowadays is Paul Nation. However, for decades researchers have been increasingly interested in clarifying what it means to know vocabulary, and thus, many researchers before him have adopted this framework, proposing various classifications of dimensions comprised in vocabulary knowledge.

One of the earlier descriptions of the multidimensional nature of word knowledge was proposed by Cronbach (1942). He produced the first modern list of five dimensions involved in vocabulary knowledge: *generalisation* (ability to define a word), *application* (ability to use a word), *breadth of meaning* (ability to know the different meanings of a word), *precision of meaning* (ability to use the different meanings appropriately in different contexts) and *availability* (ability to use a word productively). This framework does not specify the word form, but contemplates polysemy for the first time.

As time passed, research identified more components, such as orthographic knowledge, pronunciation, register, associations or morphological information. In this scenario, Jack Richards (1976) developed a broader list of what it means to know a word, including eight assumptions:

1. The native speaker of a language continues to expand his vocabulary in adulthood, whereas there is comparatively little development of syntax in adult life.
2. Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words, we also know the sort of words most likely to be found associated with the word.
3. Knowing a word implies knowing the limitations imposed on the use of the word according to variations of function and situation.
4. Knowing a word means knowing the syntactic behaviour associated with the word.
5. Knowing a word entails knowledge of the underlying form of a word and the derivatives that can be made from it.
6. Knowing a word entails knowledge of the network of associations between that word and the other words in the language.
7. Knowing a word means knowing the semantic value of a word.
8. Knowing a word means knowing many of the different meanings associated with a word.

(Richards, 1976, p.83).

Richards's (1976) list was the first exhaustive attempt to explicitly describe the most important aspects of word knowledge. In sum, he implies that knowing a word involves knowing its spoken and written form (assumption 2), knowing one or several of its meanings (assumptions 7 and 8), the associations (6), derivatives (5) and collocations of words (4 and 2), as well as connotations and restrictions on use (2 and

3). However, his description is still too vague and disorganised, as he includes assumptions on the development of general language (see number 1) which do not have to do with the vocabulary dimensions. Moreover, Richards does not specify how these dimensions of vocabulary knowledge can be identified and/or operationalised. Yet, his list was the inspiration for other scholars to try and describe the word knowledge components. One of those influenced authors was Blum-Kulka (1981), who suggested four dimensions to fully know a word: *semantic mapping* (linking words and their referents), *morpho-semantic restrictions* (accurate morphological and grammatical use of words), *collocational restrictions* (appropriateness of collocations), and *communicative functions* (pragmatic appropriacy). Later, Alexander (1982) expanded that list suggesting nine aspects that could influence the behaviour of individual words in the mental lexicon:

1. Phonological links
2. Morphological links
3. Syntactic links
4. Paradigmatic sense relationships
5. Collocational patterns (syntagmatic relationships)
6. Style and register factors
7. Knowledge of fixed expressions and idiomatic phrases
8. Connotative meaning
9. Allusional meaning (concepts and referents)

However, it was not until 1990 when a more detailed and exhaustive list of word knowledge dimensions emerged. Nation (1990) further developed Richard's list, refining and clarifying the aspects previously proposed, and even including some more, such as the receptive and productive mastery of each aspect. With its revision in 2001, Nation produces a list of nine different components, each broken into receptive [R] and productive [P] mastery (Table 1).

Table 1 Nation's (2013, p. 49) framework of the components involved in knowing a word.

FORM	Spoken	[R]	What does the word sound like?
		[P]	How is the word pronounced?
	Written	[R]	What does the word look like?
		[P]	How is the word written and spelled?
	Word parts	[R]	What parts are recognisable in this word?
		[P]	What word parts are needed to express the meaning?
MEANING	Form and meaning	[R]	What meaning does this word form signal?
		[P]	What word form can be used to express this meaning?
	Concept and referents	[R]	What is included in the concept?
		[P]	What items can the concept refer to?
	Associations	[R]	What other words does this make us think of?
		[P]	What other words could we use instead of this one?
USE	Grammatical functions	[R]	In what patterns does the word occur?
		[P]	In what patterns must we use this word?
	Collocations	[R]	What words or types of words occur with this one?
		[P]	What words or types of words must we use with this one?
	Constraints on use	[R]	Where, when and how often would we expect to meet this word?
		[P]	Where, when, and how often can we use this word?

In his framework, Nation classifies the vocabulary components under three sub-constructs, namely knowledge of form, meaning and use. Under the sub-construct of *form*, he includes not only knowledge of the spoken and written form of a word, but also the inflections and derivations of words. Under *meaning*, along the simple form–meaning mapping, also the different meanings a word can adopt and its associations are specified. Finally, the *use* category refers to the syntactic patterns that a word follows, the typical words that occur next to it and the situations and frequency with which a word can be employed.

Yet, the extensiveness of Richard's and Nation's lists began being seen as impractical, and thus some researchers encouraged the simplification of these vocabulary descriptions in an attempt to gain practicality without losing information. For example, Chapelle (1998) proposed that vocabulary should comprise four dimensions: *vocabulary size*; *knowledge of words' linguistic characteristics* (including phonemic, graphemic, morphemic, syntactic, semantic and collocational); *vocabulary organisation* (how words are represented and connected in the mental lexicon); and *lexical recognition* (processes of lexical access). Qian (1999, 2002) developed a framework in which knowing vocabulary involved interconnected information about four aspects: *vocabulary size*, *depth of vocabulary knowledge* (including pronunciation,

spelling, meaning, morphological and syntactic knowledge, collocational properties and pragmatic information), *lexical organisation* (referring to storage, representation and coaction of words), and *automaticity of receptive-productive knowledge* (all the processes involved in accessing word knowledge for receptive and productive use, which in turn involves depth information). Finally, Henriksen (1999) suggests three dimensions involved in lexical competence: *partial-precise knowledge* (the different degrees of vocabulary knowledge), *depth of knowledge* (including morphological and syntactical restrictions, paradigmatic and syntagmatic relations, and a word's referential meaning), and *receptive to productive knowledge* (ability to use words in comprehension and production).

As it can be seen from the descriptions, these so-called simplified frameworks seem to have suffered from an oversimplification, to the extent that they lack precision in their descriptions. First of all, they appear to have included multiple aspects from Nation's model into one single dimension (e.g., Qian's and Henriksen's 'depth of vocabulary knowledge' or Chapelle's 'word's linguistic characteristics'). This does not make the construct easier to interpret. Secondly, they do not state how the proposed dimensions can be operationalised, probably due to being over-crowded with several aspects. Hence, these apparent simplifications do not reflect an easier operationalisation of the construct, and consequently, most researchers seem to rely on the precise description of Nation's word knowledge conceptualisation.

In general, Nation's (2013) model of word knowledge is the most comprehensive and accurate framework of vocabulary dimensions to date, which has converted it into the conceptualisation referred to by most current vocabulary scholars. However, this taxonomy is not without complications. First of all, vocabulary knowledge is not a feature of individual words, but the whole lexical networks of participants (Meara, 1997; Meara & Wolter, 2004; Schmitt, 2014). Therefore, as Nation himself posits, it seems logical to believe that these aspects are interconnected and represent links within and between words in the mental lexicon, rather than being independent from each other. However, Nation's list, although exhaustive, does not specify the relationships between the components. This leaves important questions unanswered, such as the degree to which these separate components interrelate with each other (e.g., does knowledge of more meanings lead to better collocation knowledge?), whether some components tend to be acquired before others (e.g. are a word's derivative forms

typically learned before its collocations?) and the relative contribution of the various components to the vocabulary knowledge construct (e.g., does the form–meaning link explain most of the variance in vocabulary?).

The list's very breadth is also a disadvantage, in that it is very complex to apply in research and teaching (i.e., it is virtually impossible to measure everything). For this reason, most vocabulary research has typically treated and studied the components individually, in particular the form–meaning link. Although there is no doubt that the form–meaning link is the most essential aspect, all components need to be studied in combination to better understand vocabulary overall.

Finally, Nation himself acknowledges the utopia of intending to know all words in as much depth as the list describes, either in a L1 or L2. Yet, he believes that gaining knowledge of even some of these aspects is still a great contribution to the depth of knowledge of a word, allowing for a more appropriate and accurate use of that word. Therefore, despite its problems, Nation's list provides the best description of vocabulary knowledge to date both for teachers and researchers.

2.1.2.2. Developmental approach

This approach originated in an attempt to produce a more practical conceptualisation of vocabulary knowledge than the components approach. It considers that word knowledge exists along a continuum, from no knowledge at all, to full proficiency (Paribakht & Wesche, 1993; Read, 1997, 2000; Wesche & Paribakht, 1996). The term *developmental* refers to the use of scales for the categorisation of the degree of vocabulary knowledge. According to this approach, word knowledge is acquired in a series of different learning stages that represent each type of word knowledge, and thus it can be organised according to scales ranging from lower degrees of knowledge to more advanced degrees.

Scales have been widely used in L1 and L2 vocabulary research, and are based on the notion that certain facets of vocabulary knowledge develop before others. For example, it seems obvious to believe that a word can be recognised before it can be produced (Melka, 1997), or that some degree of vocabulary size is needed before vocabulary depth can evolve. Therefore, this developmental framework relies on, and can be applied to, different vocabulary dimensions. In this sense, it is clear that the

components approach and the developmental approach, although often treated as two separate models and described in different ways, are related.

The first developmental scale to provide an estimation of learners' word knowledge was proposed by Dale (1965) for L1 learners, and required learners to rate their vocabulary knowledge by placing a set of words onto one of four incrementally difficult categories:

> Stage 1: I never saw the word before
> Stage 2: I've heard the word but I don't know what it means
> Stage 3: I recognise the word in context, it has something to do with _____
> Stage 4: I know the word
>
> (Dale, 1965, p. 898)

In this model it is implied that the ability to recognise the word's written/spoken form develops before its meaning and other aspects, although research has not yet demonstrated whether this is the sequence in which word knowledge is indeed acquired and connected in the lexicon.

In 1997, Schmitt and Meara created a new version of Dale's (1965) scale to be used with L2 learners of English. The innovation of this scale is that it targeted English words, but was written in the participants' mother tongue. The English translation would be:

> Stage 1: I don't know this word
> Stage 2: I think I might have some sense of what this word means
> Stage 3: I think I know this word's meaning, but I am not sure
> Stage 4: I know this word
>
> (Schmitt & Meara, 1997, p. 22)

Schmitt and Meara's scale does not require any demonstration or production of knowledge, and simplifies knowledge of a word to knowing its meaning.

The scale that has been most widely employed in L2 vocabulary research is Paribakht and Wesche's (1993) Vocabulary Knowledge Scale (VKS). They expanded on Dale's developmental model by adding five instead of four stages of knowledge, and directed it to L2 learners.

> Stage 1: The word is not familiar at all
> Stage 2: The word is familiar but the meaning is not known
> Stage 3: A correct synonym or translation is given
> Stage 4: The word is used with semantic appropriateness in a sentence
> Stage 5: The word is used with semantic appropriateness and grammatical accuracy.
>
> (Paribakht & Wesche, 1993, p. 180)

As opposed to Dale's and Schmitt and Meara's model, the VKS goes beyond the knowledge of form and meaning and its link, and includes the selection of an appropriate meaning, grammatical patterns, collocations, and association. For this reason, the authors claim that the VKS can inform about breadth and depth of a learner's knowledge (Wesche & Paribakht, 1996). The scale implies that the word form develops before its meaning, and suggests that learning how to use a word in context appears later in the learning process. Also, it combines self-reported knowledge (stages 1 and 2) and productive demonstration of knowledge (stages 3, 4 and 5). However, this scale does not place much attention to receptive word knowledge, which, combined with its requirement to produce multiple aspects of depth of knowledge (i.e., associates, synonyms, and semantic and grammatical accuracy), has led researchers to argue that the scale benefits more advanced learners against beginner learners (Schmitt, 2014). Despite this, the VKS is the most famous representation of the developmental approach in L2 vocabulary research.

Wesche and Paribakht (1996) advocate the use of developmental scales such as the VKS to measure different levels of breadth and depth of word knowledge, ranging from "complete unfamiliarity, through recognition of the word and some idea of its meaning, to the ability to use the word with grammatical and semantic accuracy in a sentence" (p. 29). However, although developmental scales are a convenient (i.e., easy to produce and administer) tool to provide a general idea of depth of knowledge of a word, they have a number of limitations that has led most researchers to follow the components approach. The main issue is that they are self-reporting measures, and thus it is not clear whether the learner actually knows what they claim to know. Moreover, when producing a sentence where the correct use of a word is necessary, this can be done convincingly even without actually knowing the meaning of the target word (Schmitt, 1997). As a consequence, the scale cannot be considered a rigorous test of either vocabulary recognition or production. Moreover, the scales do not measure the different levels of sensitivity equally; that is, the different stages are not equivalent in

distance. Questions jump directly from self-report of some kind of partial knowledge to demonstration of productive skills. Therefore, the difference in the difficulty of stages is large. For example, the transition from Stage 1 to Stage 2 is smaller than from Stage 3 to Stage 4. More importantly, these developmental scales are unable to explain how and when the different aspects of knowledge are acquired. Finally, research has not yet provided empirical support that the knowledge of a word develops in the order these scales pose.

Consequently, although some studies have employed developmental scales (Folse, 2006; Kim, 2008; Yuksel & Tanriverdi, 2009), the aforementioned issues make the developmental approach problematic to be used in vocabulary research as an estimation of knowledge. In order for scales to be confidently used in research they require more precise description and empirical support. In this sense, the components approach has provided researchers with a more detailed and robust framework to assess knowledge of the different vocabulary aspects.

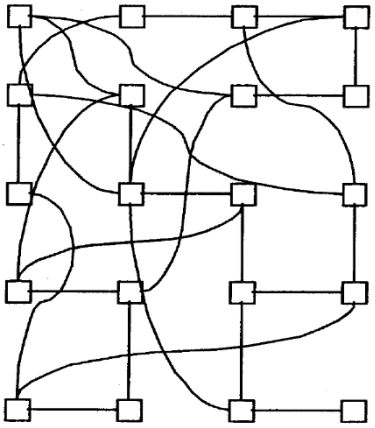
2.1.2.3. Lexical network approach

As the developmental models, this approach also emerged as a simplification of the dimensional frameworks, since some scholars considered that adding more aspects to the construct of word knowledge just made the notion more difficult to understand (Meara, 1996). The lexical networks approach views words as interconnected in the mind, so that the more and more typical links one is able to access, the greater the vocabulary knowledge, and as a result, the greater the lexical network structure (Aitchison, 2003; Meara, 1996; Meara & Wolter, 2004). Thus, they generally consider vocabulary a combination of both size and some kind of organisational connections.

The different frameworks within this approach usually describe the vocabulary construct in an abstract and almost metaphorical manner, which led Milton and Fitzpatrick (2014) to coin this approach the *metaphorical approach*. For these frameworks, word knowledge is such a complex notion that it can only be described as networks or webs of entangled and interconnected words. Included in this approach would be Aitchison's (2003) ‘web’ of words or Meara and Wolter's (2004) lexical organisation, which symbolise depth of knowledge and the interaction and connections between words as a web (see Figure 1). According to these models, lexical items will

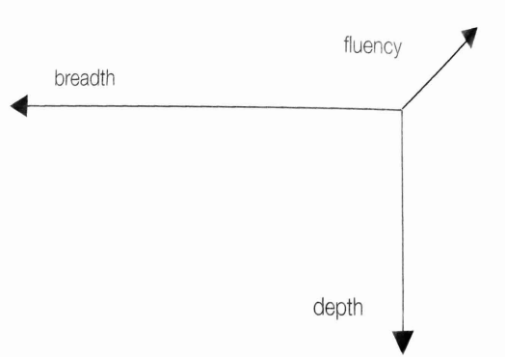
only be mastered productively if they develop the correct links and networks in the mental lexicon (Meara, 1997). Although some researchers have attempted to develop this framework into measurement models of vocabulary knowledge that tests the learners' networks between words (Meara & Wolter, 2004; Wilks & Meara, 2007), this approach generally presents vocabulary as something imperceptible, and thus it is considered impractical for vocabulary research and teaching.

Figure 1 A network model of vocabulary (Meara & Wolter, 2004, p. 89).



Another symbolic explanation of the word knowledge construct that can arguably be considered a lexical network description is Daller, Milton, and Treffers-Daller's (2007) lexical space. This lexical space involves the facets of breadth and depth, to which fluency is added. For them, this fluency refers to how automatically and fluently learners can access the links between words in the mental lexicon and use them. According to Milton and Fitzpatrick (2014), this lexical space metaphor can prove useful in certain cases, for example to distinguish between learners that know a large number of words but have difficulties to access them in the lexical network, and thus, use them, and those whose vocabulary size is small but can access the words fluently and use them naturally and appropriately. Nevertheless, the fact that this description represents the three facets of breadth, depth and fluency as developing independently from each other (see Figure 2) seems counterintuitive and unlikely (see Qian, 2002; Schmitt, 1998, 2014; Tannenbaum et al., 2006), and thus it has been claimed that this framework cannot be considered a realistic representation of word knowledge.

Figure 2 The lexical space (Daller et al., 2007, p. 8).



Overall, these lexical network frameworks are not sufficiently concrete and practical to be able to thoroughly describe word knowledge or explain how words are acquired and linked with each other. For these reasons, although theoretically interesting, the lexical networks approach has not been considered useful for vocabulary acquisition research.

2.1.2.4. Conceptual approaches to vocabulary knowledge and vocabulary acquisition theories

The various approaches to vocabulary knowledge are related to different views of how vocabulary is acquired. The components descriptions of vocabulary knowledge are directly connected to Perfetti's (1992, 2007) *Lexical Quality Hypothesis*, which claims that high-quality vocabulary representations are needed for skilled language use. According to Perfetti (2007) the words a person knows vary widely in depth or lexical quality, so that those of high quality include phonological, orthographical, grammatical and semantic links, while those of lower quality lack some of this information or the bonds between them. Therefore, lexical knowledge or depth of knowledge would involve knowledge of a word's form, meaning and use. The lexical quality hypothesis, thus, suggests that obtaining multiple types of word knowledge is the key to master language use and comprehension. This hypothesis also implies that the acquisition of words is not a fixed process, but develops over time through extensive exposure to the target language, involving changes in the lexical system.

Following this lexical quality hypothesis, it has been suggested that when we start to learn a language only a limited number of words can be recognised, and only for the core meaning of those words. Then, we continue learning more and more words,

constituting breadth of vocabulary. Only with exposure and experience to the language can those words be developed further and defined with more detail, progressing from only recognition to the ability to access and produce those words. Finally, links start being created with other words in various contexts and becoming more fluent, which in turn supports the learning of new words (Beck et al., 2013; Li & Kirby, 2015; Perfetti, 2007; Tannenbaum et al., 2006). Since this hypothesis claims that more exposure to a word will improve its depth and quality, it can be considered as related to the linguistic usage-based approaches, in which language is conceptualised as a system determined by exposure and experience (Bybee, 2006). Although this lexical quality hypothesis has obvious connections with the components approach, it is also related to the lexical network approaches, as it considers depth of vocabulary knowledge to further develop by building networks between words and fluency (Daller et al., 2007; Haastrup & Henriksen, 2000; Meara, 1996).

The developmental approach, however, is more directly connected to the *Processability Theory*. Pienemann (1998, 2005) suggested this theory to try to explain linguistic acquisition orders in terms of a set of processing routines that occur at different points of development. He describes that the acquisition of language production follows some basic procedures, from knowing a word or lemma, to being able to identify its category, to producing noun and verb phrases, and then sentences with correct word order, to finally being able to produce subordinate clauses. Fundamentally, this theory is based on the notion that these processing procedures are implicationally ordered, that is, hierarchical and mastered in a sequence. Nevertheless, although this theory was specifically designed as a model of L2 language processing, it has been suggested to be applicable as a theory of language acquisition, since language acquisition is constrained by language processing (Ellis, 2008). In this sense, the acquisition of lexis and other language aspects would follow developmental sequences in which failure to master the easier stages impedes the access to the more difficult ones that depend on them. Research based on this theory examines the sequence of acquisition of certain linguistic features by individual learners. Related to this hierarchical theory, Fitzpatrick and Clenton (2017) propose a model which follows a developmental scale to interpret the learner's performance on vocabulary productive tests with what learners can do in that language. This model is called the Vocabulary Test Capture Model, and implies that productive knowledge progresses from being able

to produce the word form, to then use the words for appropriate referents or L1 words, later use them with semantic appropriateness in context, and finally use them with semantic appropriateness and grammatical accuracy in context. According to this model, the different levels of productive vocabulary knowledge are meant to follow an implicational scale, suggesting that they develop hierarchically and are mastered in a sequence. However, the validity of these implicational scales, and thus the processability theory, for vocabulary acquisition still need to be examined. (More substantive discussion on this developmental method of analysis will be reviewed in Chapter 4).

Although different theories exist, it is still unclear how vocabulary is stored and processed in the mental lexicon (i.e., the mental dictionary where humans store the words they have some knowledge of). It is known that words are not unrelated and independent from each other, but rather they are highly organised and linked in multiple ways to the rest of the words stored in the lexicon, so that learning one lexical item has some effect on learning others (Meara & Wolter, 2004). Therefore, to develop full knowledge of a word it is necessary to build a rich and densely interrelated mental lexicon, which favours more rapid, comprehensive, and accurate networks between words (Cremer, Dingshoff, de Beer, & Schoonen, 2010). However, examining the links among words is proving a very complex and challenging task.

It seems that, as in the L1, L2 learners develop their mental lexicon by adding and reorganising the connections between words. Williams and Cheung (2011), in a semantic priming study on L1 Chinese learners of French, found that newly learned words did not simply adopt the L1 meanings. Rather, the new words automatically acquired their own semantic representations, which were associated with the contexts and meaning situations in which the words were learned. For example, when encountering a new word in a L2 (e.g., *écureuil* in French) and learning that its equivalent in the L1 is *squirrel*, one would expect that if the student knows that the word in the L1 (*squirrel*) is semantically associated with the word *nut*, then the new L2 word *écureuil* would also be. However, these authors found that the new word created its own associations based on the context in which it was learnt (e.g., a bushy tail based on a fairy tale character), and not on the meaning of the L1 word.

Therefore, L2 vocabulary learning is not seen as simply the integration of new knowledge into the existing L1 system, but as establishing connections between aspects

of word knowledge through exposure to the word in varying contexts. That is, as suggested by the lexical quality hypothesis, word knowledge develops from experiences and encounters with the language and connections between previous word knowledge and the new information, which will develop further with more and more varied exposures to a word (Perfetti, Wlotko, & Hart, 2005).

Overall, it is generally accepted that vocabulary acquisition is incremental in nature, progressing along different stages until complete mastery. However, since word knowledge is acquired through multiple and varied language experiences (e.g., through both explicit instruction and incidental exposure: Schmitt, 2008), the acquisition of words is not a fixed process. Rather, research has also found that vocabulary learning varies from item to item and stage to stage, so that each lexical unit goes through different stages at different times and for different amounts of time until all the word knowledge aspects needed to employ a word accurately in different situations are acquired (Fitzpatrick, 2012; Henriksen, 1999). Moreover, if language is shaped by the experiences to that language, vocabulary knowledge would be subject to individual variations. Consequently, each learner is likely to have somewhat idiosyncratic lexicons based on their individual experiences with language. This changeable nature has led researchers to suggest that the mental lexicon is a dynamic system (Churchill, 2008), which makes it even more difficult to examine the links between words in the mental lexicon , and is one reason for the lack of a generally accepted theory of how the mental lexicon functions and vocabulary is acquired.

2.1.3. Summary of the conceptualisation issues

Vocabulary knowledge is a complex construct, and as a result, it has been conceptualised in multiple and diverse ways. Nevertheless, from the review above it seems clear that the various frameworks coincide in describing vocabulary as a multidimensional construct, in which quality or depth of vocabulary knowledge plays a central role. In fact, these different frameworks are not mutually exclusive, as they share some common aspects in a lesser or higher degree depending on their description (Schmitt, 2014). In spite of these commonalities, research has typically favoured the use of the components approach, for providing a more precise understanding of the word knowledge construct as a whole. Hence, this components approach, in particular

Nation's (2013) model of vocabulary knowledge, is currently the preferred framework to conceptualise vocabulary in research (Cheng & Matthews, 2018; Li & Kirby, 2015; Pigada & Schmitt, 2006; Van Zeeland & Schmitt, 2013; Webb, 2005, 2007a).

Yet, none of the word knowledge frameworks described above have been empirically examined, and thus, supported. Consequently, the actual nature of vocabulary knowledge and how it develops is not yet clear, and needs to be examined (Li & Kirby, 2015; Milton & Fitzpatrick, 2014).

2.2. Vocabulary Knowledge and SLA

2.2.1. Nation's word knowledge components: what do we know so far?

The above section showed that, although there is no accepted universal framework of vocabulary knowledge, Nation's (2013) model seems to represent the most authoritative taxonomy currently available (Milton, 2013; Schmitt, 2014). His comprehensive framework has been established for over two decades as a prompt that knowing a word involves much more than only mapping form and meaning. Therefore, despite having been the main focus of vocabulary research for decades, researchers now agree that simply measuring the form–meaning link of words (i.e., size) results in incomplete information and an overgeneralisation of vocabulary knowledge. For this reason, in recent years research has explored the behaviour of the different components of word knowledge to try and obtain a better understanding of the vocabulary knowledge construct. In the following sections, I will describe the different components of word knowledge individually, summarising the research that has been carried out on them, and therefore, what we know about these aspects to date.

2.2.1.1. Spoken form

Nation (2013) describes knowing a word's spoken form as being able to recognise it when heard and produce it in speech. The field of vocabulary research has generally disregarded the study of spoken vocabulary knowledge in favour of written aspects. As a result, far less is known about the acquisition of the spoken form of words as compared to the written form.

In general, research has shown that knowledge of the spoken form of a word poses more problems to L2 learners than its written form (Cheng & Matthews, 2018; Milton & Hopkins, 2006; Milton & Riordan, 2006; Shimamoto, 2000). For example, measuring knowledge of written form recall, written form recognition and spoken form recall by EFL Chinese learners, Cheng and Matthews (2018) found that the spoken form recall was the most complex aspect (48% mean score), being worse known than the written form recognition (82%) and recall (55%). Similarly, Shimamoto (2000) found that Japanese university students were able to translate more L2 words into the L1 when they were presented in the written form (65.2%) than in the spoken form (56.6%).

It has been suggested that one reason for this generally better knowledge of the written form over the spoken form of a word is their different acquisition situations. Acquiring the spoken form is more difficult because it occurs under the pressure of real time, which hinders the learning. The general recognition and production of a word's form involves the simultaneous activation of several competing word neighbours (i.e., words that differ from each other only in one grapheme or phoneme) (Diependaele, Lemhöfer, & Brysbaert, 2013). Unlike in reading or writing, a listener/speaker has to select the appropriate receptive candidate in a limited amount of time, and pronounce its different phonemes accurately and uninterruptedly, making fluency an essential factor in spoken form recognition and production. This required fluency is the reason that L2 learners usually fail to recognise words in continuous speech that they are able to recognise in isolation (Van Zeeland & Schmitt, 2013a). Another characteristic that makes spoken form knowledge more difficult than that of its written form is the great variety of ways in which a word can be pronounced, due to the adjacent words in speech as well as dialectal (e.g., American vs. British English), regional (e.g., Midlands vs. London) or social (e.g., RP vs. slang) features (Nation, 2013). This variation is not as pronounced in the written form, making it easier for L2 learners to acquire.

Although research has typically found that the knowledge of a word' spoken form is worse than that of its written form knowledge, some scholars suggest that it depends on the learners' vocabulary size. Milton and Hopkins (2006) administered a yes/no spoken form recognition test (A_Lex) and its written version (X_Lex (Meara & Milton, 2003)) to Greek and Arabic students, and found that, while the word's written form knowledge of the Greek learners was better than their spoken form, the Arabic learners had slightly better knowledge of spoken form than written. They concluded

that with larger L2 vocabulary sizes, the written form recognition knowledge is better than the spoken form, while this pattern reverses when the vocabulary size is small (these results were also duplicated by Milton and Riordan (2006) for Farsi EFL learners). These studies also indicate, unsurprisingly, that the spoken form knowledge is greater in L2 learners of higher proficiencies. Milton and Hopkins (2006) found that the spoken form knowledge increased as the Greek's participants' proficiency level was higher, and thus, advanced learners seem able to recognise many more spoken forms of words than lower proficiency learners.

Finally, research seems to indicate that learners need a smaller spoken form size to understand spoken discourse than they need written forms to comprehend written discourse. It has been argued that L2 learners need a vocabulary of between 2,000-3,000 word families in English (providing 95% coverage) in order to communicate orally in basic, everyday informal situations (Nation, 2006; Van Zeeland & Schmitt, 2013), while the written form size requirement to enable initial engagement with reading texts at the same basic level would be of 4,000-5,000 families (95% coverage) (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

Despite the difference in knowledge between a word's spoken and written forms by L2 learners, research has found moderate to large correlations between them. For example, Milton and Hopkins (2006) found a strong correlation of .68 between the A_Lex and X_Lex, and Cheng and Matthews (2018) correlations of between .59 and .81 between two written form and a spoken form tests. These results suggest that the spoken and written forms of a word are interrelated aspects, although the different measurement instruments and participants' proficiencies might yield slightly different results.

2.2.1.2. Written form

A word's written form is fully mastered when learners are able to recognise it correctly and fluently when reading and able to produce it in writing. Research suggests that the written form of words (i.e., spelling) is likely to be the first word knowledge aspect to be acquired by L2 learners. Schmitt (1998) found that students mastered the spelling of words before derivational forms and meanings. Measuring more word knowledge components, Webb (2005, 2007a) found that the written form was the aspect

that experienced the highest gains, and that this written form advantage was consistent for receptive and productive knowledge across all encounters (1, 3, 7 and 10 repetitions), although it was always greater in receptive knowledge. This finding was also supported by Chen and Truscott (2010), Pellicer-Sánchez and Schmitt (2010), and Pigada and Schmitt (2006).

Knowledge of the written form of words has also been found to be learnt incrementally, and not always in a forward manner. For example, Fitzpatrick (2012) tracked longitudinally the acquisition of various word knowledge aspects by a Chinese learner of English, and found that the acquisition of written productive form went through three stages: approximation to the right form, knowledge of the correct form but produced inconsistently at different points in time, to finally production of the written form consistently. The inconsistent development may respond to the formal and semantic factors that have been found to affect a word's written form knowledge, such as the word length, morphological structure, concreteness or word frequency (De Groot, Borgwaldt, Bos, & Van den Eijnden, 2002; Zoccolotti, De Luca, Di Filippo, Judica, & Martelli, 2009). These factors do not only affect the recognition of a word's written form, but also how accurately and quickly it is acquired and processed.

As with any aspect, the recognition of the written form has been more commonly researched than its production, primarily with the aim of assessing learners' receptive vocabulary size. Although vocabulary size is typically operationalised as knowledge of the form–meaning link, some scholars have employed tests measuring the written form recognition of words as proxies of receptive vocabulary size (e.g., Milton & Hopkins, 2006; Milton & Riordan, 2006; Pellicer-Sánchez & Schmitt, 2012). Generally, these tests require learners reporting whether a word was known or not (e.g., X_Lex (Meara & Milton, 2003), Y_Lex (Meara & Miralpeix, 2006)), or whether it exists in the specific language or not while controlling the reaction time (e.g., LexTALE (Lemhöfer & Broersma, 2012), VORST (Laufer & Nation, 2001)). Research employing these tests has found that receptive knowledge of the written form correlates with performance in literacy skills, such as reading and listening (e.g., Milton, 2013; Van Zeeland, 2014).

The recognition of written form has also been employed to measure how quickly words are accessed in the mental lexicon. Research in psycholinguistics has shown that L2 learners usually require more time to recognise a word's written form than native speakers, with higher-proficiency learners accessing them quicker than low-proficiency

learners (e.g., Conklin & Schmitt, 2008). Also, L2 learners seem to process the written form of unknown L2 words in an attention-demanding manner, but this process becomes more automatic, and thus more native-like, as learners become more familiar with those words (Pellicer-Sánchez, 2016). In general, the speed of recognition of the written form is related to the development of fluency of vocabulary knowledge, and hence can provide an indication of this development (Schmitt, 2010; Segalowitz, Segalowitz, & Wood, 1998).

2.2.1.3. Word parts and derivational morphology

The word parts aspect refers to the fact that words can be morphologically complex, comprising affixes that are added to the stem form and can change its grammatical category. This aspect is related to the notion of morphological awareness, defined as "the ability to reflect on and manipulate morphemes and word formation rules in a language", with morphemes defined as the smallest meaningful unit within a word (Kuo & Anderson, 2006, p. 161).

In the field of L2 vocabulary research, the morphological aspect that has been of most interest in L2 vocabulary research is affixation, in particular suffixation. Research has found that the acquisition of inflectional suffixes seems to be easier for EFL learners than that of derivational suffixes, probably due to the fact that inflectional suffixes are rule-based, explicitly taught and very frequent in English[3] (Reynolds, 2015; Schmitt & Meara, 1997; Ward & Chuenjundaeng, 2009). For example, Schmitt and Meara (1997) tested EFL Japanese learners' verb suffix recognition and recall, and found that their participants had a generally poor knowledge of suffixes, except the regular inflectional ones (*-ing, -ed, -s*). In a recent study, Reynolds (2015) examined the incidental learning of L2 words that appeared in a text with the same form, inflections or derivations of those words. He found that words that did not vary in form or only in their inflections were known better than words that varied in their derivations, confirming the general disadvantage of derivational knowledge over inflectional.

Probably due to the difficulty of derivational morphology for L2 learners, research within the study of L2 suffixation has been particularly interested in this aspect.

[3] The easier acquisition of inflectional suffixes over derivational ones may not hold when learning morphologically more complex languages, such as French, German or Spanish (Laufer, 1997).

Knowledge of derivational morphology has been divided into three aspects: relational, syntactic and distributional knowledge. According to Kuo and Anderson (2006), *relational knowledge* is the ability to recognise the stem in morphologically complex words and understand the relationship between stem and suffix. For example, recognising that the stem of 'productive' is 'product'. *Syntactic knowledge* is the understanding of how a derivational suffix changes the part of speech of a word and how those derivatives work in context. That is, knowing that 'productive' is a verb and 'production' is a noun and that they are related but need to be used differently in sentences. Finally, *distributional knowledge* is the ability to understand how suffixes are constrained by the syntactic category of the stem, that is, which suffixes can apply to a noun (e.g., *-tion*), or an adjective (e.g., *-able*). These types of derivational knowledge are acquired in a developmental sequence, where relational knowledge develops before syntactic mastery, with distributional knowledge being the most complex aspect to be acquired even by L1 learners (McCutchen, Green, & Abbott, 2008).

In the L2 vocabulary field, research on this word parts aspect has mainly focused on assessing syntactic derivational knowledge, by measuring knowledge of the part of speech, and distributional derivational knowledge. Regarding the part of speech, research has typically focused on measuring four of the multiple types of word class, namely nouns, verbs, adjectives and adverbs. From these, nouns have been found to be generally the easiest to learn, probably due to their greater imageability, while adverbs are considered the most difficult word class (Dóczi & Kormos, 2016; Ellis & Beaton, 1993; Schmitt, 1998; Schmitt & Zimmerman, 2002).

As for distributional derivational knowledge (generally labelled 'derivational knowledge' in L2 vocabulary research), the findings indicate that this aspect is very challenging for even advanced L2 learners, particularly in its productive mastery (e.g., Saigh, 2015; Schmitt, 1998; Schmitt & Zimmerman, 2002). Schmitt (1998) studied productive knowledge of derivations in very advanced learners of English, and found that they only had partial derivational knowledge, particularly of nouns and verbs. Schmitt and Zimmerman (2002) analysed the productive knowledge of derivatives in the four main word classes: noun, verb, adjective and adverb, and found that participants usually had partial knowledge of the derivatives (usually two or three word class derivative forms) but that it was difficult for participants to know which suffix should

be applied to which word. These findings suggest that distributional derivational knowledge seems to be acquired incrementally, and that knowing a word family member does not imply knowledge of the rest of the word family members.

Finally, it has been found that general derivational knowledge is not mastered by native speakers until late in life (Nagy, Diakidoy, & Anderson, 1993), and thus it is to be expected that non-native speakers would have problems acquiring this knowledge and that it would appear later than other aspects (Jiang, 2000; Schmitt, 2014). The difficulty for L2 learners to develop derivational competence is thought to be due to factors such as the suffixes' regularity and predictability of form and function, their frequency, their productivity, and their transparency (Bauer & Nation, 1993; Nagy et al., 1993; Nation, 2013; Thornbury, 2002).

Despite being a challenge for L2 learners, awareness of word formation directly benefits learners' general vocabulary knowledge. It is believed that knowledge of affixes and roots can help students understand how words are formed, and this morphological information can behave as the bond between word form (written and spoken) and word meaning knowledge (Kieffer & Lesaux, 2008; Perfetti, 2007; Proctor et al., 2009; Qian, 1999). Moreover, derivational awareness has been found to be a predictor of literary skills such as spelling, listening and reading (Fracasso, Bangs, & Binder, 2016), as well as being related to the learners' vocabulary size (Kieffer & Lesaux, 2012a; Lo, Anderson, & Bunch-Crump, 2017; Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997). For these reasons, knowledge of derivatives has been found to be a valuable source for L2 learners to quickly increase not only their vocabulary size, but also their depth of knowledge receptively and productively (Nagy et al., 1993; Schmitt, 2010). Therefore, it has been suggested that knowledge of affixation and derivational patterns should be an essential aspect of explicit instruction from beginning stages until advanced levels (Carlisle, 2010; McCutchen & Stull, 2015; Nagy et al., 1993; Reynolds, 2015; Schmitt & Zimmerman, 2002).

2.2.1.4. Form and meaning

This aspect of word knowledge concerns the link between a word's form and its meaning. Receptive knowledge of the form–meaning link involves mapping a given written or spoken L2 form to its meaning, while productive knowledge involves

mapping an existing meaning to its appropriate written or spoken L2 form. This link is considered essential in communication and, therefore, some authors argue that it is the first of the word knowledge aspects to be acquired (Agustín Llach & Moreno Espinosa, 2014; Laufer & Goldstein, 2004).

This aspect is by far the most extensively researched vocabulary component, partly because of its importance for communication, partly due to its relative simplicity in understanding and measurement. In fact, it has been contended that most vocabulary measures used in research are assessing productive and/or receptive knowledge of the form–meaning link to some extent, as even those tests designed to measure other aspects (such as word association tests) require implicit knowledge of the form–meaning link (Laufer et al., 2004; Laufer & Goldstein, 2004).

The form–meaning link aspect has been described as comprising four types of knowledge within it, namely form recall, form recognition, meaning recall and meaning recognition (in Laufer and Goldstein's (2004) terms: active recall, active recognition, passive recall and passive recognition). Applying implicational scaling analysis, Laufer and Goldstein found that these types of form–meaning link knowledge follow a consistent order of difficulty. According to this order, learners would first recognise the word's meaning, then the word's form, then they would recall the word's meaning, and, in the final stage of the acquisition of this link, they would be able to recall a word's form. This finding is compatible with the evidence that learners have a larger receptive form–meaning knowledge than productive, and that the acquisition of the form–meaning link improves incrementally with time, exposure and practice (Fitzpatrick, 2012; Laufer et al., 2004; Laufer & Paribakht, 1998).

Although the form–meaning link is generally considered an easy aspect, its acquisition is not as simple as it may seem, especially when the word's form is complex. Indeed, in studies comparing acquisition of multiple aspects of non-words and very low-frequency words, Webb (2007a) and Chen and Truscott (2010) have found that the form–meaning link is the last aspect to be mastered. Moreover, it has been discovered that in learning circumstances when: 1) the form is irregular orthographically or phonetically, 2) in which one form does not only encode on meaning, but multiple meanings depending on the context, or 3) when multiple forms encode the same meaning, mastering the form–meaning link is more difficult, particularly for lower-proficiency learners (VanPatten, Williams, & Rott, 2004). For these reasons, the

relationship between the form and meaning is not completely straight-forward, and there is still some debate whether the form is learnt before the meaning, or vice versa (Fitzpatrick, 2012; Webb, 2005; Wolter, 2009).

As mentioned above, knowledge of the form–meaning link is commonly employed as an indication of the vocabulary size of a person. Although there are different possible ways of measuring the form–meaning link (see Section 2.2.5), the most widely employed tests to assess its receptive mastery are the Vocabulary Size Test (VST) (Nation & Beglar, 2007) and the Vocabulary Levels Test (VLT) (Schmitt et al., 2001). The VST has been largely used to measure vocabulary size, although it has been recently found that it tends to overestimate learners' knowledge of the form–meaning link of words (Gyllstad, Vilkaitė, & Schmitt, 2015). The VLT, although originally designed to provide an estimation of the learners' vocabulary profile (Kremmel & Schmitt, 2017), has also been used as an indication of vocabulary size, insofar that it measures the form–meaning link (Nation, 1990). The productive vocabulary size is generally measured with the productive version of the VLT, the PVLT (Laufer & Nation, 1995), which requires learners to recall the form of a word to complete a short gapped sentence. In general, research has found that the productive knowledge of form–meaning (i.e., form recall, PVLT) is lower than the receptive form–meaning link knowledge (i.e., form recognition, VLT), although they correlate strongly (e.g., Laufer & Paribakht, 1998; Nemati, 2010).

2.2.1.5. Concept and referents

As described by Nation (2013), a word's concepts and referents involves knowing what is included in it (i.e., abstract knowledge of the meaning of that word) and what it can refer to (i.e., the concrete examples represented by the word). According to the *Revised Hierarchical Model* (Kroll, Michael, Tokowicz, & Dufour, 2002), learners acquire knowledge of the words' semantic information in the L2 through linking those words to the knowledge they already have in the L1 translation equivalents, and then adding the L2 labels to them. That is, while L1 children learn a word's meaning(s) and form together, L2 learners seem to know the meaning(s) of a word first through their L1 and then learn its form. However, meaning is a very complex construct, as the relationship between words and what they represent is arbitrary (i.e.,

conventionalised by a language community). Words can be polysemous (i.e., have more than one meaning), homonymous (i.e., have different meanings from different origins), they can be concrete (i.e., it can be experienced by the senses) or abstract, and they often have both literal and idiomatic meanings. Moreover, learners can encounter words in the L2 whose concepts and referents are unknown in the L1. All these aspects of a word's meaning can affect its acquisition (Laufer, 1997; Webb & Nation, 2017). Indeed, Williams and Cheung (2011) found that, just as L1 children, L2 learners do not simply learn words by adopting the L1 meanings, but that they are able to create specific semantic associations for the L2 forms based on the context in which they are acquired. As a result, the claim that meaning appears first is not as straightforward.

The knowledge of concept and referents, although largely overlooked in vocabulary studies, has primarily been conceptualised in research as knowledge of multiple meanings of a word (Li & Kirby, 2015; Schmitt, 1998; Tannenbaum et al., 2006). When learning words with multiple meanings, it seems that the core (and generally most frequent) sense of a word is acquired before the less frequent senses (Verspoor & Lowie, 2003), and sometimes is the only sense known by learners (August, Carlo, Dressler, & Snow, 2005). It has also been found that acquiring multiple meanings of words is very challenging even for advanced learners. Schmitt (1998) studied the knowledge of different meanings of words over an academic year and found that, although their knowledge of different senses improved, none of the learners acquired full knowledge of all the meanings. Similarly, Tannenbaum et al. (2006) examined the knowledge of multiple meanings of words by 8-9-year-old native English children and found that they knew only 67% of the multiple meanings of the words. Therefore, this aspect is considered difficult for learners, and thus believed to appear later in life (Wolter, 2009).

Research has also discovered that the acquisition of the multiple meanings of a word is not a fixed state, but a gradual process that develops over a long period of time (Henriksen, 1999; Schmitt, 1998). Salsbury, Crossley, & McNamara (2011) examined the free production of polysemous words by beginner learners over one year, and found that, although they produced polysemous words from the beginning, only during the second and third trimester had their knowledge of the different meanings of a word started to increase. It has been claimed that the reason for multiple meanings to occur late is because it involves learning different senses of a word, and also understanding

the semantic networks between words in the mental lexicon. As Cremer et al. (2010) put it, "words get their meaning from their relations to other words". Therefore, the more words' meanings we acquire, that is, the more links between words we have, the deeper our word knowledge is (Aitchison, 2003; Henriksen, 1999). In fact, it has been found that new words in another language acquire their own semantic representations by associations with the aspects of meaning that are active at the time they were learnt (Williams & Cheung, 2011). Therefore, semantic knowledge of L2 words develops by establishing connections between aspects of word knowledge through experience, and thus, it can only be acquired through multiple and varied language encounters (Cremer et al., 2010; Perfetti, 2007).

2.2.1.6. Associations

Nation (2013) describes knowledge of word associations as knowing as many words connected to the target one as possible, and being able to use appropriately all those words instead of the target one. This aspect includes both knowledge of formally associated words (e.g., homophones or homographs), as well as the semantic relations between words (e.g., paradigmatic or syntagmatic associations).

Research commonly assesses this type of knowledge by means of free word association tasks, which are claimed to examine the organisation and links between lexical items in the mental lexicon (Cremer et al., 2010). In these tasks, participants are given a set of target words, one by one, and they have to respond to each of them with the first word that comes to their minds. An established test of free productive associations is the Lex30 (Meara & Fitzpatrick, 2000), which assesses the form recall of associations presented in isolation. Association tasks can, therefore, provide an insight into how learners' mental lexicons develop and are organised, as well as finding similarities and differences in the individuals' mental lexicons. Therefore, as claimed by Fitzpatrick and Munby (2014), research based on word associations is directly related to the lexical network conceptualisations of word knowledge (e.g., Aitchison, 2003; Meara, 1996).

Despite the agreement that the analysis of word association patterns can provide information about the way the mental lexicon functions, it is still unclear what kind of word knowledge information word association tasks provide. Some authors claim that

word associations inform us about depth of vocabulary knowledge (Read, 1993, 1997), while others consider that they measure both size and depth of word knowledge, as well as (and even mainly) the structure and relationships between words in the learners' mental lexicon (Cremer et al., 2010; Schmitt, 1998; Wolter, 2002). This difference means that the interpretation of word association tasks generally depends on the particular purpose of the research.

In L2 vocabulary research, word association tasks have been used mainly to assess depth of vocabulary knowledge, generally analysing the free answers by either response type or response item. Research characterising the responses in types, typically distinguish between *syntagmatic* (e.g., collocations), *paradigmatic* (e.g., synonyms, superordinates) and *clang* (e.g., neighbours or homonyms) associations. Some examples of this approach are Cremer et al. (2010), Fitzpatrick (2006), and Nissen and Henriksen (2006), who compare the pattern of word association responses between L1 and L2 learners. In this research, the form-based associations are considered reflections of poorer associations in the lexicon, while meaning-based, in particular paradigmatic associations, are deemed indicators of good links between words (Cremer et al., 2010). Research describing the response item focuses mainly on the strength of links in the lexicon between certain words (e.g., binomials *left* > *right*, *salt* > *pepper*). In this approach, the answers of L2 learners are compared to the native-like association response patterns to determine the L2 learners' proficiency (the more native-like answers are, the more proficient learners are considered). Examples of this approach are Schmitt (1998) and Wolter (2002).

In general, research in L2 word associations has found that the learning of associations is a slow process, and that even proficient L2 learners provide a smaller number and more varied responses to word association tasks than L1 learners (Cremer et al., 2010; Fitzpatrick, 2012; Haastrup & Henriksen, 2000; Nissen & Henriksen, 2006). Research has also suggested that L2 learners' word association responses become more native-like, and thus their association knowledge is better, as proficiency increases (Dóczi & Kormos, 2016; Fitzpatrick & Munby, 2014; Nurweni & Read, 1999; Qian, 1999, 2002; Schmitt & Meara, 1997; Verhallen & Schoonen, 1993). However, the opposite has also been claimed, where stronger correlations have been found between word association knowledge and size in learners with smaller vocabulary sizes (Henriksen, 2008; Noro, 2002). Therefore, the general agreement in L2 word

association research is that the association behaviour of L2 learners deviates from that of L1 learners due to multiple factors, such as proficiency, instruction or cultural differences. Furthermore, the responses to these free association tasks are based on subconscious, subjective connections, and thus these links between words constantly change (Cremer et al., 2010). Consequently, the results of the free association tasks only reflect the lexical connections at a specific point in time, not for all word knowledge. For those reasons, it has been argued that learners' word knowledge should be measured with controlled tasks, such as employing word definition activities or multiple-choice tests (Dóczi & Kormos, 2016; Qian, 1999, 2002, Read, 1993, 1998). This claim led to the development of the most widely used association tasks in L2 research: Read's (1993, 1998) Word Association Test (WAT), which measures semantic and collocational associations between words, and Qian's (1999, 2002) revised version of Read's WAT, named Depth of Vocabulary Knowledge Measure (DVK), measuring knowledge of collocations and meaning, including polysemy and synonyms. Still, these tests suffer from ambiguity about which type of knowledge the results actually represent, and thus researchers claim that the different aspects should be better assessed concurrently but with individual tests.

2.2.1.7. Grammatical functions

Knowing the grammatical functions of a word refers to knowing the patterns in which that word occurs and can be employed. This knowledge comprises different types of words' patterns, for example, the purely grammatical patterns of words, such as a verb's transitivity or intransitivity, or the lexicogrammatical patterns, such as knowing whether a verb is associated with the pattern *to + infinitive* or the *-ing* construction (Hoey, 2005; Sinclair, 1991).

The investigation of language patterns in the mental lexicon has been of interest for decades (Wray, 2002, 2008). Already in 1954, Hornby identified linguistic patterns for verbs, nouns, and adjectives. The use of electronic corpora has helped facilitate this task, providing strong support for the importance of patterns in language and the inseparability of grammar and vocabulary knowledge within them (Granger, 1998; Nattinger & DeCarrico, 1992; Sinclair, 1991). In fact, more advanced corpora, and

software for their analysis, have contributed to an increasing attention in studying lexicogrammatical patterns in L2 language.

However, the great difficulty of describing specifically the grammatical patterns of language use means that this type of constructions has not been thoroughly explored, particularly in L2 vocabulary acquisition research. As a result, it is still uncertain how this grammatical functions aspect of word knowledge is learnt, although it is considered to be challenging for L2 learners. One study that examined this aspect in L2 learners is Pigada and Schmitt (2006). They measured productive knowledge of grammatical functions by requiring L2 learners to report the prepositions that followed specific verbs, along with spelling recall and meaning recall knowledge of those verbs, and found that grammatical patterns was the most complex knowledge type for learners to acquire. Grammatical functions have also been found to be related to vocabulary size, so that the larger the vocabulary proficiency of learners, the more complex grammatical patterns they know (David, Myles, Rogers, & Rule, 2009).

Furthermore, the intrinsic difficulty of defining, identifying and operationalising grammatical patterns in language has led to confusion in the study of this aspect. Some vocabulary studies have investigated grammatical functions, but understood them as knowledge of the words' part of speech, instead of the learners' knowledge of the grammatical behaviour of words. For example, Webb (2005, 2007a) and Chen and Truscott (2010) studied the acquisition of what they called productive and receptive 'grammatical functions' by asking L2 EFL learners to recognise and produce a word in the correct word class. Consequently, little is still known about the knowledge of actual grammatical functions by L2 learners. However, Nation (2013) argues that knowing these grammatical patterns determines how well a word can be used in language, and therefore more research is needed to better understand this aspect in L2 vocabulary acquisition.

2.2.1.8. Collocations

Although they have been defined in multiple ways, collocations can be understood in general terms as combinations of words that co-occur in the input. They have traditionally been classified into two main categories: grammatical or syntactic collocations (i.e., those that include prepositions, e.g. *immune to*, *tolerance for*), and

semantic or lexical collocations (i.e., combinations of noun, verbs, adjectives and adverbs, e.g., *soft noise*, *unforeseen circumstances*) (Bahns, 1993; Benson, Benson, & Ilson, 1997; Henriksen, 2013). According to Nation (2013), knowing collocations refers to understanding which words occur and must be employed with a target lexical item. Thus, this aspect of word knowledge is directly related to other components of Nation's framework, namely associations and grammatical functions. Regarding the former, word association research commonly measures collocations as the best example of syntagmatic associations (e.g., WAT or DVK). As for the latter, collocations are part of the lexicogrammatical patterns of a word, although it could be argued that the grammatical functions aspect is more concerned with the study of grammatical collocations and the aspect labelled as 'collocations' more related to research of lexical collocations. Therefore, knowledge of collocations also comprises, to some degree, knowledge of association and grammatical functions.

While grammatical patterns have not been extensively researched in L2 vocabulary acquisition, lexical patterns in the form of lexical collocations have been a popular phenomenon in L2 language acquisition research since Firth (1957) first spoke about these constructions. Nesselhauf (2003, 2005) differentiates two traditions in research on collocations: the *phraseological approach*, which is based on the meaning of collocations (e.g., Benson et al., 1997; Cowie, 1994; Hornby, 1954), and the *frequency based approach*, which focuses on words' number and strength of co-occurrence (e.g., Hoey, 2005; Sinclair, 1991). Recent work on collocations seems to suggest that the frequency-based approach is more common nowadays, although the phraseological approach is still used. In fact, these approaches are not seen as mutually exclusive, and several studies combine both methods to provide a more comprehensive description of collocations (see Barfield & Gyllstad, 2009; Davies & Gardner, 2010). In these studies, the researchers use frequency-based techniques to produce a list of collocations and then apply a phraseological approach to select the final target collocations and interpret results. Furthermore, the phraseological approach has suggested that collocations should be considered as independent lexical units, instead of an aspect of word knowledge. In this sense, the other types of word knowledge described by Nation would apply to knowledge of collocations. Some researchers have adopted this approach, measuring multiple types of knowledge of collocations, such as

the written form and the form–meaning link both receptively and productively (e.g., Webb, Newton, & Chang, 2013).

Collocations (as a type of formulaic sequences) are ubiquitous in language. Yet, a number of studies have shown that they are difficult to acquire by L2 learners, their knowledge is limited and their development tends to be slow even for high proficiency students (Laufer & Waldman, 2011; Li & Schmitt, 2010). Most of the vocabulary research has been based on lexical collocations, as they seem easier to identify and test, and appear more salient for L2 speakers. Research suggests that learners do not generally have problems recognising collocations, as long as they know the individual words that form them (Laufer & Waldman, 2011). Rather, their main difficulties seem to be related to the productive and appropriate use of those collocations, where even advanced learners produce collocation errors even when there are no other problems with the use of words (Levitzky-Aviad & Laufer, 2013; McGee, 2009; Siyanova & Schmitt, 2008). Moreover, researchers seem to agree that the L1 influence is the main source of problems for learners when they try to use collocations, as they often directly transfer the collocational patterns from the L1 to the L2 (Bahns, 1993; Nesselhauf, 2003, 2005; Wolter & Gyllstad, 2011). It has also been shown that collocational knowledge is strongly related to vocabulary proficiency, with lower-proficiency learners performing worse in collocation knowledge. For example, Gyllstad (2009) found that the vocabulary size of highly-proficient L2 English students correlated at the .90 level with two measures of collocational knowledge (a multiple-choice form recognition test and a yes/no collocation judgement test).

Overall, research suggests that, although collocations are difficult for L2 learners, they do know and make use of these formulas. Yet, learners rely on a small set of favourite collocations that they tend to overuse, underuse or misuse (Durrant & Schmitt, 2009; Granger, 1998; Kim & Kim, 2012; Li & Schmitt, 2010). Furthermore, they may also make use of strategies like synonymy or paraphrasing to compensate for the lack of collocational knowledge (Farghal & Obiedat, 1995), which has led to the belief that, even for advanced speakers, collocational knowledge falls behind their general vocabulary knowledge (Bahns & Eldaw, 1993; Irujo, 1993). Finally, given the benefits that the use of formulaic language has on the accuracy and fluency of communication, collocations have been claimed to be a target for vocabulary instruction

(Boers & Lindstromberg, 2009; Coxhead, 2008; Durrant, 2009; Simpson-Vlach & Ellis, 2010).

2.2.1.9. Constraints on use (register, frequency and fluency)

Constraints on use is probably the aspect of Nation's framework that has received the least attention in vocabulary knowledge research. As a result, it is unclear how knowledge of the constraints on the use of words is acquired. A possible reason for research overlooking this aspect is its intrinsic complexity, which has led to an ambiguous and imprecise description of what the constraints on use component involves as compared to other aspects. Nation explicitly cites frequency and register as features involved in the constraints on the use of words, and also implies that other elements exist within this aspect, for example fluency. Yet, no clear definition is provided by Nation for any of these features, making them elusive concepts in research. In this section, I will summarise the knowledge of register, frequency and fluency in relation to vocabulary knowledge.

Register is an important constraint on the use of words, but it is also a very complex feature to delimit and describe. Halliday (1994) was one of the first linguists interested in defining this concept, and described it as the lexical and grammatical choices made by speakers depending on the context, the interlocutors or the purpose. Thus, a words' register constraints on use would involve knowing which lexical candidates are more appropriate in a specific situation. The register information can be temporal (old-fashioned vs. modern words), geographical (dialects) or social (slang, formal), and should vary based on the content, aim, relationships between interlocutors or mode of discourse (Halliday, 1994). Thus, register would include, for example, information about whether a word is from American or British English, whether it is formal or informal, polite or impolite, or typical from written or spoken discourse. However, deciding on the specific register information that a word holds is not always easy, and different researchers/speakers actually disagree on, for example, the formality or politeness of certain words. In addition, register is a dynamic concept, and can change in relatively short time within a language community, which has led some authors to claim that register should not be described in a dichotomous manner for learners. Nevertheless, although register has been a focus of interest in areas such as pragmatics

and sociolinguistics, it has not been thoroughly researched in the field of vocabulary studies, and hence little is known about how it is acquired by L2 learners.

Frequency is another characteristic of a word's constraints on use. Unlike register, the concept of frequency is easier to capture and thus has been extensively studied in vocabulary research. It is now agreed that frequency of occurrence of a word affects the acquisition, use and processing of that word (Ellis, 2002). In fact, it is recognised as one of the best predictors of usefulness and acquisition of lexical items (Leech, Rayson, & Wilson, 2001; Schmitt & Schmitt, 2014; Webb & Nation, 2017). By definition, the most frequent items in language are more likely to occur in input, and thus learners would acquire them before less frequent items (Schmitt, 2010). Indeed, research shows that native speakers have relatively good intuitions about the actual frequency of words, even though they do not always match corpora frequency data (Alderson, 2007; McGee, 2008). In L2 vocabulary, research has found that non-native speakers also have intuitions of the frequency of L2 words, although their estimates are less accurate than those of native speakers (Schmitt & Dunham, 1999; Siyanova & Schmitt, 2008; Siyanova-Chanturia & Spina, 2015). Generally, frequency is commonly treated in L2 research as a factor affecting language (see Section 2.2.4), rather than as part of a word knowledge component.

Finally, fluency has been considered another feature that constrains the use of a word. It is widely accepted in L2 research that being able to process and use language quickly, fluently and accurately is an indicator of advanced language proficiency. Fluency can be understood as the speed of recall and recognition of words in a language, and develops with the enhancement of older links in the mental lexicon and the development of new links from practice and experience that can make the structure of the lexicon more efficient (Segalowitz & Segalowitz, 1993). Fluency is also linked to the degree to which vocabulary knowledge can be processed with automaticity (Segalowitz & Hulstijn, 2005), that is, can be undertaken with accuracy, speed, and effortlessness, spending very few cognitive resources (Fraser, 2007). With more practice and familiarity with a word, learners become more proficient and fluent, and their language processes move from being controlled and reliant on mental capacity to being automatic (Hulstijn, Van Gelderen, & Schoonen, 2009; Segalowitz & Hulstijn, 2005). Thus, in order to develop fluency and automaticity of words, the acquisition of different aspects of word knowledge is essential. The importance of fluency for

receptive and productive language proficiency has led researchers to not only consider it part of depth of vocabulary (e.g., Segalowitz & Hulstijn, 2005; Segalowitz et al., 1998), but also an independent component of word knowledge that adds to size and depth (Daller et al., 2007; Meara, 1996; Tannenbaum et al., 2006). However, this aspect is very difficult to develop, although research suggests that it is related to vocabulary size, so that the fluency of lexical access speed increases as the vocabulary size and meaning of words grows (Beck et al., 2013; Laufer & Nation, 2001). Yet, research in the fluency of access of vocabulary in L2 is scarce, and thus our understanding of this component is limited.

Overall, developing knowledge of constraints on use seems to be challenging for L2 learners. Since these constraints vary depending on language and cultural differences, it has been suggested that L2 learners would benefit from explicit instruction on these constraints on use information, enabling them to use words appropriately and accurately in different contexts (Nation, 2013; Wilks, 2014). Nevertheless, in order to further investigate this component and obtain a better understanding of how it functions in the L2 learners' minds, the first step research needs to take is proposing more precise and elaborated definitions of what constraints on use comprises.

2.2.2. How do word knowledge components relate to each other?

The general main focus of vocabulary research on measuring the form–meaning link of words has two major consequences for our knowledge of the vocabulary construct: an oversimplification of what vocabulary knowledge involves, and a comparatively limited knowledge of the other components contained in vocabulary knowledge. The first matter responds to the fact that most of the conclusions drawn on vocabulary teaching, learning, and processing rely primarily on the study of just one aspect of vocabulary. Therefore, it is unknown whether these conclusions can be applied to each word knowledge aspect and to vocabulary overall. The latter issue implies that there is only partial knowledge of the rest of the components involved in depth of vocabulary (Qian, 2002; Li & Kirby, 2015), and thus the interrelationships between these components seem unclear: does improvement in one aspect lead to improvement in others?

The implementation of the components approach, which divides the word knowledge construct into smaller units each time, has resulted in research concentrating on smaller aspects of vocabulary, making it more difficult to obtain a big picture of how vocabulary in its multiple dimensions is acquired (Milton & Fitzpatrick, 2014). In fact, most vocabulary research has typically studied the components individually, compared at most with vocabulary size, e.g. form (Meara & Milton, 2003; Milton & Hopkins, 2006); the form–meaning link (Agustín Llach & Moreno Espinosa, 2014; Laufer & Goldstein, 2004); derivatives (Kieffer & Lesaux, 2012b; Schmitt & Zimmerman, 2002); collocations (Gyllstad & Wolter, 2015; Peters, 2016; Siyanova-Chanturia & Spina, 2015); polysemy (Khodadady & Khaghaninizhad, 2012; Verspoor & Lowie, 2003); and associations (Fitzpatrick & Izura, 2011; Qian, 2002; Read, 1998).

Furthermore, Nation's (2013) description of word knowledge dimensions in the form of a list can give the erroneous, but unintended, impression that these components are independent from each other. Indeed, most researchers, including Nation himself, argue that those dimensions are intrinsically interconnected and interact with each other in vocabulary use (Chui, 2006; Qian, 2002; Schmitt, 2014; Webb, 2005). For example, strong correlations have been found between knowledge of the spoken and written form of words by Milton and Hopkins (2006) (r = .68), Van Zeeland (2014) (r = .85), or Cheng and Matthews (2018) (r = .59-.81). Similarly, correlations have been found between most aspects and vocabulary size (i.e., form–meaning link), such as derivatives (Mochizuki & Aizawa, 2000), associations (Qian, 1999, 2002), or collocations (Gyllstad, 2009).

Other research has attempted to suggest when specific components would be acquired as compared to others. For example, the acquisition of the form–meaning link has been claimed to be possibly the first and most central aspect of vocabulary acquisition (Agustín Llach & Moreno Espinosa, 2014; Cheng & Matthews, 2018; Henriksen, 1999; Laufer et al., 2004). Cheng and Matthews (2018) actually state that "[i]n the earliest stages of learning a word, a language learner must be able to link the form of the word with the word's meaning" (p. 2). Within this form–meaning aspect, receptive knowledge has been found to occur earlier than productive knowledge of either form or meaning alone (Laufer & Goldstein, 2004). Moreover, studies suggests that knowledge of derivatives (Jiang, 2000; Nagy et al., 1993) and knowledge of the meanings of words (Wolter, 2009) would progress slowly and could potentially be the

last stages in L2 word knowledge. However, these conclusions on the order of acquisition of word knowledge have been based on studies assessing the components individually, instead of comparing the knowledge of multiple aspects simultaneously.

Although these single-component studies are informative for the component in question, they provide only limited insight into vocabulary knowledge as a whole, and how the various components fit together (Li & Kirby, 2015). Exploring the construct of vocabulary knowledge with a components approach requires breaking the overall construct into units that are more easily accessible and measured, but still obtaining information about the full knowledge of a word (Schmitt, 1997). Hence, this has led a number of scholars (Henriksen, 1999; Li & Kirby, 2015; Milton & Fitzpatrick, 2014; Read, 2000; Schmitt, 2010; Webb, 2005) to encourage the measurement of multiple components concurrently (including their receptive and productive masteries) with a battery of tests as the only way to fully explore the multidimensionality of vocabulary knowledge. These different tests would tap into different word knowledge cognitive processes at different sensitivities. However, given the complexity involved, there have been relatively few scholars who have taken up this challenge. Their studies are reviewed in the next section.

2.2.3. Multicomponent vocabulary research

One of the first scholars to explore knowledge of multiple word knowledge components concurrently was Schmitt (1998). He conducted a longitudinal study of the vocabulary knowledge of four ESL learners in university settings. He measured the productive knowledge of four word knowledge components of 11 words: written form, derivatives knowledge, associations, and knowledge of multiple meanings. He found that 1) when knowledge of meaning increased, the associations and derivatives scores also increased proportionally; 2) as associations scores improved, so did the derivatives and meaning scores; and 3) when derivatives knowledge improved, associations and meaning also increased proportionally. Therefore, he concluded that these four components of word knowledge were interrelated somehow, and improved more or less in a parallel manner.

Chui (2006) studied the knowledge of vocabulary of 186 ESL learners. Twenty words from the AWL were tested across four components: word class recognition,

meaning recall, collocation recognition, and derivative form production. Although the participants were university students, she found that general knowledge of the AWL words was relatively modest: word class 87.6%, meaning 64.5%, collocation 57.4%, and derivatives 52.3%. These results suggest that word class recognition and meaning recall were better known by the students, and therefore might be acquired earlier than the knowledge of the productive derivative form or the recognition of collocations.

Shimamoto (2000) measured four word knowledge aspects of a large number of words (50 words): meaning recall from spoken stimulus, meaning recall from a written stimulus, recognition of paradigmatic associations and collocation recognition. He found that the recognition aspects were better known than the recall aspects, with knowledge of paradigmatic associations being higher than collocations knowledge, and written meaning recall higher than the spoken meaning recall.

While the above three studies explored the extant *knowledge* of components, Webb (2005, 2007a) studied the *acquisition* of vocabulary, examining how 66 and 121 (respectively) Japanese EFL students acquired 10-20 nonwords from various exposures. He used an extensive battery of tests to measure five components of word knowledge, both productively and receptively: orthography (written form), form–meaning link, syntax (syntagmatic associations), grammatical functions (word class) and (paradigmatic) associations. He found that orthography was the component that enjoyed the highest gains, both receptively and productively, at all levels of exposure. The rest of the components varied over the different exposures to the words, but there was a tendency for syntax and the form–meaning link to obtain lower scores. Overall, he found that the more exposures to a word, the greater the gains in all the different components, and showed that the different components developed in a parallel manner, but at slightly different rates.

Following Webb (2007a), Chen and Truscott (2010) investigated the acquisition of four word knowledge components by 72 Taiwanese university students (orthography, parts of speech, and associations (both receptively and productively), and form–meaning link (receptively only)), and found that increasing repetitions lead to better knowledge in all the different components, although the gains in knowledge varied depending on each component. The results showed that receptive knowledge of each component was always higher than their productive counterpart, and that the form–meaning receptive aspect required the greatest number of exposures to be learnt.

Pellicer-Sánchez & Schmitt (2010) also examined the incidental acquisition of spelling recognition, meaning recognition, word class recall and meaning recall in 20 Spanish EFL learners. They found that, after more than 10 exposures, the recognition aspects were known much better than the recall aspects, indicating that they are easier to acquire.

These multicomponent studies suggest that the various components are interrelated, with growth in one leading to growth in others, but that some appear to be acquired before others. For example, recognition of spelling seems to be an aspect generally achieved earlier. However, a summary of the results of previous studies that looks at their tentative acquisition patterns (see Appendix 1 for this summary) shows that there is simply not enough evidence to indicate an acquisition order of these components, or to specify the nature of their interrelationships. This summary also shows that the complexity in studying vocabulary knowledge not only resides in the multiple aspects involved, but also in the changing, incremental nature of each component. The degree of (gained) knowledge of the various components was different across studies, even when the exact same measures were used (Webb 2005, 2007a, 2007b, 2009). Thus, the results are inconsistent regarding what appears to be acquisition orders based on the mean scores. This is partly because word knowledge components have been conceptualised and tested inconsistently (see Section 2.2.5). Also, researchers have applied different learning techniques (e.g., Pellicer-Sánchez & Schmitt, 2010; Webb, 2005, 2007b), which could have enhanced specific types of word knowledge (Webb, 2005, 2009), or no learning technique at all (e.g., Chui, 2006; Schmitt, 1998). However, the main reason for the inconsistency in results is that these studies were not designed with the purpose of exploring an acquisition order in mind, and only examined the strength of the relationships between word knowledge components using simple correlational and mean scores analyses. These mean scores can be misleading, as individual learners do not necessarily experience the same order of difficulty as the sample as a whole. In order to explore the possibility of an acquisition order of word knowledge components, specific statistical analyses are required, and the design of the study needs to accommodate to the features of such analyses. (A broader discussion of these key methodological issues is included in Chapter 4, referring to Implicational Scaling Analysis, and in Chapter 6, referring to Structural Equation Modelling). Although there are some studies that have employed more refined statistical

methods for the examination of vocabulary knowledge, such as structural equation modelling or linear regressions (e.g., Kieffer & Lesaux, 2008, 2012a, 2012b; Kremmel, Brunfaut, & Alderson, 2017; Li & Kirby, 2015; Ma, 2015; Shiotsu & Weir, 2007; Tannenbaum et al., 2006; Tseng & Schmitt, 2008; Zhang, 2012; Zhang & Koda, 2011), their main aim was to investigate how lexical knowledge influences other aspects of language learning (e.g., reading comprehension or language motivation). Therefore, these studies, cannot inform either on the relationships between word knowledge components, or between the various components and overall vocabulary knowledge.

In sum, although the above studies describe some types of word knowledge, they have limitations which restrict their ability to inform how multiple knowledge components of words develop in relation to each other. Some of these limitations are: not measuring multiple word knowledge components individually (Kieffer & Lesaux, 2008, 2012c; Kremmel et al., 2017; Shiotsu & Weir, 2007; Tannenbaum et al., 2006; Zhang, 2012; Zhang & Koda, 2011); testing only receptive or productive mastery (Schmitt, 1998; Tannenbaum et al., 2006), or testing both but inconsistently across components (Chen & Truscott, 2010; Chui, 2006; Li & Kirby, 2015); measuring multiple components, but for different target words (Qian, 2002; Kieffer & Lesaux, 2012a; Li & Kirby, 2015; Milton & Hopkins, 2006); relatively few words tested (Chen & Truscott, 2010; Schmitt, 1998; Shiotsu & Weir, 2007; Webb, 2005); interpretation of mean scores and correlations only (Chen & Truscott, 2010; Chui, 2006; Webb, 2005, 2007a); low numbers of participants (Schmitt, 1998; Webb, 2005); and use of nonwords (Webb, 2005, 2007a).

These limitations are at the heart of the current lack of accepted theory of how vocabulary knowledge overall is acquired. Researchers in the field agree that in order to attempt to establish such theory it is necessary to understand how the various dimensions of vocabulary knowledge are interconnected in the mental lexicon (Milton & Fitzpatrick, 2014; Schmitt, 2010; Webb & Nation, 2017). Thus, research needs to be conducted that concurrently measures multiple word knowledge components in order to investigate the acquisition and relationships between these components and obtain a better understanding of the vocabulary knowledge construct. This would allow us to develop a multidimensional theory of vocabulary acquisition that describes the various components and processes involved in vocabulary development (Henriksen, 1999; Schmitt, 2014).

2.2.4. Factors influencing vocabulary acquisition

The acquisition of vocabulary is not only governed by the specific aspect of word knowledge being learnt. Learning words has also been shown to be affected by a variety of factors that may increase or decrease their learning burden. They can be classified as socio-contextual factors or linguistic factors.

2.2.4.1. Socio-contextual factors

The socio-contextual factors influencing vocabulary acquisition involve features such as the learners' age, motivation, the way in which a word is taught, or the type of input to the word.

Research has shown that the age at which words are acquired affects the knowledge learners have of them, even over linguistic aspects such as word frequency, word length or word neighbourhood density (Garlock, Walley, & Metsala, 2001; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). This age effect on vocabulary acquisition, however, seems to somewhat differ from the acquisition of other language elements such as phonology, in that research has found that later-starting children can have better general vocabulary knowledge than earlier-starting learners (Miralpeix, 2006).

The learner's motivation also has a direct effect on vocabulary learning and the strategies employed for their acquisition (Ma, 2015; Tseng & Schmitt, 2008; Tseng, Dörnyei, & Schmitt, 2006). A clear example of the learners' motivation influence on lexical acquisition is the general interest in learning taboo or swear words when acquiring a second language, regardless of their intrinsic difficulty or usefulness.

The learning context is another factor to account for in vocabulary acquisition. Whether a particular word is learnt under deliberate or incidental learning conditions can also affect its difficulty. Although researchers agree that lexical items can be learnt incidentally by L2 learners (Laufer & Rozovski-Roitblat, 2011; Pellicer-Sánchez & Schmitt, 2010; Van Zeeland & Schmitt, 2013; Webb, 2005; Webb et al., 2013), some types of words and word knowledge aspects may be more amenable to deliberate learning than incidental learning. For example, the morphological rules and formal patterns of a linguistic system are more apt to be conveyed by deliberate teaching. Explicit instruction on these formal systems of a language will allow learners to gain

knowledge of these features of words more rapidly, helping reduce the learning difficulty they pose on their acquisition (Dóczi & Kormos, 2016; Webb & Nation, 2017). Conversely, polysemous words comprise different contextual information within their different meanings, and thus would require multiple and varied encounters to be learnt. This enhanced input is easier to be provided by incidental learning through, for example, extensive reading, than by the limited time of deliberate learning. Therefore, the vocabulary learning situation needs to be adapted to the particular type of words and word aspects in order to facilitate their acquisition.

Also, the manner in which lexical items are taught can affect their learning burden. For example, teaching two new words together that have formal or semantic similarities (i.e., opposites or near synonyms) can lead to cross association, and make them more difficult to be learnt (e.g., teaching *left* and *right*, or *latter* and *former* together) (Schmitt, 2010). Similarly, learning words in semantically related sets, such as parts of the house, days of the week or colours, have been found to make their acquisition 50-100% more difficult than learning unrelated words (Nation, 2000). Hence, the selection of words presented together to learners have an effect on their acquisition, and thus should be taken into account in teaching.

Despite their importance in vocabulary acquisition, these socio-contextual features of word learning difficulty are variable factors and generally more difficult to control by teachers and learners than linguistic-related factors. Therefore, research on vocabulary acquisition has more commonly focused on examining the linguistic factors that affect vocabulary learning.

2.2.4.2. Linguistic factors

The linguistic-related factors in vocabulary acquisition comprise features derived from the linguistic systems of the L1 and the L2, the differences and similarities between them, the interaction between words within the L2, and the intrinsic characteristics of the specific words. The latter are generally referred to as *intralexical factors*, whilst all the rest belong to the *interlexical factors*. Both types have been found to affect the learning burden of words (see Laufer, 1997; Nation, 2013; Schmitt, 2010). Unlike the socio-cultural features, these linguistic-related factors are generally fixed for

particular languages and words, and thus have been more thoroughly explored in vocabulary research.

One of the aspects that most affects the learning burden of a L2 word is how well it fits into the L1 system. That is, how the L2 matches the L1 lexical features. Learners rely on their L1 when they begin to learn another language, and therefore L2 words that are more similar to the L1 equivalent will be easier to learn than words that differ (De Groot, 2006). A particularly important case is that of *cognates*, that is, words that due to their common origin, are semantically and orthographically and/or phonologically similar in the L1 and L2. These lexical similarities, known as the 'cognate facilitation effect', can reduce the learning burden of L2 words (Rogers, Webb, & Nakata, 2015; Rosselli, Ardila, Jurado, & Salvatierra, 2012; Tonzar, Lotto, & Job, 2009). It has been argued that around 60% of the English words came from French, Latin or Greek (Webb & Nation, 2017), so learners whose mother tongue shares these etymological roots (such as English and Spanish), would experience this facilitation effect when learning and processing the form–meaning link of these words. Substantial research documents that bilingual learners from various mother tongues, both having received and not having received formal instruction on cognates, can use cognates as an advantage in the identification, processing and understanding of cognate English words (August et al., 2005; Carlo et al., 2004; Chen, Ramirez, Luo, Geva, & Ku, 2012; Dressler, Carlo, Snow, August, & White, 2011; Méndez Pérez, Peña, & Bedore, 2010; Ramírez, Chen, & Pasquarella, 2013).

But this L1 facilitation of word learning is not only unique to the similarities between the lexical systems of the L1 and L2. Rather, it is also influenced by the similarities between the phonological, orthographic, morphological, and syntactical systems. Therefore, differences in the alphabet, spelling, pronunciation, affixes, meanings, or syntactic patterns of a word in the L1 and L2 would largely increase the learning burden. For example, a speaker of a syllable-timed language such as Spanish would find it difficult to learn the pronunciation of a stress-timed language like English. Similarly, a learner of L2 English that comes from an alphabetic writing system language (i.e., Spanish) will have less problems with its written form than a learner from a logographic language such as Chinese (Webb & Nation, 2017). Thus, the difficulty of learning a L2 in general will be greater for L2 learners as more of these various linguistic systems differ from the L1.

In addition, the characteristics of the L2 various linguistic systems not only affect the acquisition of words in the L2, but also the relationships between those words within the L2. This way, L2 words with regular orthographic and phonological form, regular and transparent morphology, and one meaning for one form have been found to be easier to learn than those with irregular phonemes or letters, irregular and complex morphology, idiomatic meanings and with multiple meanings (Laufer, 1997; VanPatten et al., 2004). Therefore, learners need to be made aware of the similarities and differences between the L1 and the L2 systems in order to be able to adopt strategies to learn and use L2 vocabulary more effectively and efficiently (Swan, 1997).

Apart from the linguistic typicality (i.e., the degree to which the formal and semantic systems of the L2 resemble those of the L1), the difficulty of learning a word is affected by the specific features of the L2 words. These intrinsic characteristics involve the word's length or frequency, the number of senses a word can have and how they differ from each other (i.e., polysemy vs. homonymy), the literal and idiomatic meanings of a word, or the imageability and concreteness of a word's meaning (i.e., how easy it is to imagine and experience the concept by the senses (De Groot, 2006)) (Laufer, 1997).

Among these intrinsic word features, the word's frequency has been found to have a stronger effect on vocabulary acquisition, being often considered the best predictor of vocabulary knowledge. Research has shown that knowledge of individual words develops according to word frequency, so that more frequent words are more likely to be encountered in discourse repeatedly, and thus, the most likely to be acquired (Schmitt et al., 2001). This effect has been found to be robust for both L1 and L2 learners, learners from different L1s (Laufer & Paribakht, 1998), and for both individual words and formulaic language (Durrant, 2014; Schmitt & Schmitt, 2014).

Word frequency not only affects vocabulary learning directly. It is also interrelated, and sometimes inseparable, from other features of words, and thus, also affects these processes indirectly. For example, most frequent words have been found to be shorter, while less frequent words are generally longer, which makes the two factors of frequency and word length interdependent (Milton, 2009). Frequency has also been found to be related to the learners' proficiency, with the effect of frequency decreasing as language proficiency develops (Diependaele et al., 2013). Furthermore, it has been claimed that frequency is likely to be a factor affecting the acquisition of most

of the word knowledge components described by Nation (Crossley, Salsbury, & McNamara, 2010; Schmitt, 2010; Wesche & Paribakht, 1996), although this effect has been mainly demonstrated for the form–meaning link, but not most other aspects (Leech et al., 2001). Overall, all these characteristics make frequency a very important factor in the acquisition of general vocabulary.

Finally, the learners' proficiency/vocabulary size has also been found to influence vocabulary acquisition. Vocabulary size seems to affect most word knowledge components, with higher-proficiency learners performing better than lower-proficiency learners (e.g., Dóczi & Kormos, 2016; Peters, 2016). Indeed, positive and strong correlations have been found between vocabulary size and knowledge of associations (e.g., Qian, 1999, 2002; Nurweni & Read, 1999; Schmitt & Meara, 1997), or knowledge of affixes (e.g., Schmitt & Meara, 1997; Mochizuki & Aizawa, 2000). Therefore, the likelihood of a learner mastering specific words and types of word knowledge is greater as their L2 proficiency develops. (A more thorough discussion of vocabulary research on linguistic factors is provided in Section 5.1.1, Chapter 5).

To sum up, whether words are easy or difficult to learn depends on their intrinsic difficulty, the similarities with the language(s) known (particularly the L1) and the regularity of the words in the L2 system. Although they are described and examined independently, these linguistic and socio-contextual factors interact with each other in the process of acquiring vocabulary. That is, as shown by studies measuring the effect of various factors on vocabulary knowledge (Nguyen & Webb, 2017; Vidal, 2011; Wolter & Yamashita, 2017), the learning burden of a word is determined by multiple factors in combination.

2.2.5. Measuring vocabulary knowledge

The existence of different descriptions and ways of understanding vocabulary knowledge inevitably leads to various means of assessing this mastery.

As stated above, most vocabulary research has concentrated on the examination of single word knowledge components. Yet, to date there are no standardised or ideal tests that measure each of these aspects. Different types of tests with various elicitation techniques tap into different cognitive processes and degrees of knowledge, even within the same word knowledge aspect (Fitzpatrick & Clenton, 2017; Webb, 2007b). As a

result, the findings in the single-component studies differ due partly to the different measurement instruments employed. For example, the strength of correlation between knowledge of written and spoken word forms varied from .68 to .85 when they were assessed by means of yes/no form recognition tests (A_Lex and X_Lex, Milton & Hopkins (2006)) and when they were tested employing a meaning recall interview combined with a yes/no test (Van Zeeland, 2014).

The one word knowledge aspect that has experienced the development of some type of standard test is the form–meaning link, probably due to its connection with the assessment of vocabulary size and its interest in vocabulary research. Common formats for measuring the receptive form–meaning mapping, and thus provide an estimation of vocabulary size, are word translation into L1 (e.g., Takala, 1985; Webb, 2005), multiple-choice to select or match the right definition or synonym (e.g., Schmitt et al., 2001), word recognition tasks or yes/no tasks (e.g., Meara & Jones, 1990; Milton & Hopkins, 2006), or oral picture selection tests (e.g., Dunn & Dunn, 2007). The productive aspect of the form–meaning link has been less explored, but accepted tests are L1 to L2 translation (e.g., Webb, 2007a), or sentence completion exercises of form-recall (e.g., Laufer & Nation, 1999). Probably the most standard tests of the form–meaning link are the VST and the VLT (Nation, 1990) for receptive knowledge and the PVLT for productive knowledge (Laufer & Nation, 1999). Among these, the VLT and its productive version PVLT are perhaps the best known and most frequently employed tests to measure the form–meaning relationship. Both tests consist of four word frequency bands (2,000, 3,000, 5,000 and 10,000) as well as words from the AWL, for the VLT, and the University Word Level for the PVLT. The VLT measures knowledge of 30 words at each frequency level by means of matching a word with its correct definition from multiple choices, and the PVLT assesses 90 words in total by recalling the correct word form elicited in a short, gapped sentence. Although not originally designed for this purpose, these tests have been widely employed as estimations of vocabulary size in L2 learners (Meara & Fitzpatrick, 2000; Read, 2000), and as a result they have experienced multiple updates (see VST by Nation and Beglar (2007), VLT by Schmitt et al., (2001), new VLT by McLean and Kramer (2015), and updated VLT by Webb, Sasao, and Balance (2017)).

However, not all the word knowledge aspects have generated as much interest as the form–meaning link, and therefore, the test formats employed to measure other

components are far less established. In addition, the incomplete information about the vocabulary knowledge of a person provided by single vocabulary measures has led most researchers to acknowledge the importance of measuring multiple word knowledge components concurrently. Despite this agreement, there is to date a lack of standardised tests that measure these components simultaneously (Henriksen, 1999; Read, 2004; Schmitt, 2014).

Some attempts have been made to measure multiple word knowledge components in a single test. For example, Read's (1993, 1998) WAT assesses the knowledge of paradigmatic (i.e., synonyms) and syntagmatic (i.e., collocations) associations of words in combination. Similarly, Qian's (1999, 2002) DVK was especially designed to measure learners' knowledge of associations (synonyms and collocations) as well as polysemy. Although these tests were created to provide a consistent measure of depth of vocabulary knowledge (indeed they are more informative than single-component tests), they are limited in what they can say about a word. First of all, they measure only knowledge of associations and/or multiple meanings of words, thus missing other aspects involved in depth of word knowledge. Although it can arguably be claimed that knowledge of other aspects are implied within it (i.e., in order to recognise a second meaning of a word a learner is expected to be able to correctly recognise the form of the word and have established the form–meaning link already with its most common meaning), this knowledge cannot be always assumed. Consequently, these formats are not sensitive enough to measure depth of vocabulary knowledge, since it can be that the reason why a learner cannot recognise, for example, the collocations of a word is either that they know the form, and the form–meaning link of that word, but not the collocations, or that they do not recognise the form and/or form–meaning link. Therefore, with this format beginner learners will not be given any credit for their knowledge, and thus tests need to examine the fine-grained individual vocabulary aspects. Secondly, these tests cannot inform about overall word knowledge because they only include recognition of associations and/or polysemy. Receptive and productive tests measure word knowledge aspects at different sensitivity levels, and thus both are essential to obtain a more accurate and precise insight of the knowledge of vocabulary aspects (Henriksen, 1999; Nagy, Herman, & Anderson, 1985; Webb, 2005). Therefore, despite the importance of developing instruments for the assessment of different word knowledge aspects, these WAT and DKV tests cannot tap into the

whole complexity of vocabulary knowledge, and thus cannot be accepted as standardised tests of depth of word knowledge.

As a result of these issues, researchers have highlighted that in order to fully measure a learner's word knowledge and obtain a complete understanding of the vocabulary knowledge construct, these aspects need to be addressed by means of a set of individual tests that assess with precision the multiple components, both in their recall and recognition mastery (Fitzpatrick & Milton, 2014; Li & Kirby, 2015; Nation & Webb, 2011; Schmitt, 1998; Webb, 2005). These tests would independently measure the strength of knowledge in each vocabulary component, but would be connected in a battery with the rest through the set of words measured.

The multicomponent studies in Section 2.2.3 follow this approach to a certain extent. Still, as reported in Appendix 1, the attempted orders of difficulty derived from each study are inconsistent even when they measure the same word knowledge components. A reason for these inconsistencies is the diverse ways of conceptualising and testing the word knowledge components. Appendix 2 provides a summary of the tests employed in multicomponent studies for the different word knowledge components. In this appendix it can be seen that vocabulary knowledge aspects have been tested inconsistently. Different studies measure different components, or they measure the same components but with different tests. For example, knowledge of word parts has been assessed with tests of part of speech knowledge (Webb, 2007a; Chen & Truscott, 2010), inflectional and derivational suffixes (Schmitt & Meara, 1997) or only derivational suffixes (Chui, 2006), or identifying the base of multi-morphemic words (Li & Kirby, 2015). Likewise, knowledge of multiple meanings, although not commonly investigated, has been measured through word definition tasks, in which learners have to provide explanations of the meaning of the target word (Tannenbaum et al., 2006), by providing explanations to prompt words designed to elicit specific secondary meanings (Schmitt, 1998), or by identifying the sentence(s) in which a target word is used correctly (Kieffer & Lesaux, 2012a; Li & Kirby, 2015). The written word form knowledge is generally measured by selecting and recalling the right spelling of a word (Chen & Truscott, 2010; Pellicer-Sánchez & Schmitt, 2010; Webb, 2007a). However, the level of precision required in the spelling of the words (i.e., only exact or partial knowledge is accepted) yields different results (Webb, 2008). Consequently, the lack of consensus on how to best test overall vocabulary knowledge as well as specific

components makes it difficult to compare and generalise the results of these studies, limiting our knowledge about the relationships between vocabulary knowledge aspects.

To sum up, the results found in research regarding knowledge of vocabulary components vary depending on the different measures used. Although there are various testing methods for receptive skills (e.g., L2 to L1 translation, multiple-choice or matching exercises), productive skills (L1 to L2 translation, writing a sentence, description tasks), form–meaning link (e.g., yes/no test, multiple-choice), derivatives (part of speech, select the stem, select the suffix), etc., there is no ideal format to date for any of these degrees of knowledge. Therefore, standardised tests still need to be developed to assess overall vocabulary knowledge through a set of measures of the discrete word knowledge aspects but applied in an integrated manner (Cheng & Matthews, 2018; Li & Kirby, 2015; Schmitt, 2014). Thus far, researchers agree that the best form of evaluation of overall vocabulary knowledge is by separately testing different vocabulary components at different sensitivity levels with a test battery. This approach needs more attention in research for the examination of the relationships and order of acquisition of vocabulary components (refer to Chapter 3 for a more detailed discussion of vocabulary measurement).

2.2.6. Summary of the issues with L2 vocabulary knowledge

Vocabulary knowledge research moved from a core emphasis on exploring primarily the form–meaning mapping to the examination of other word knowledge components, such as the written form, derivatives, associations and collocations. This research provided essential information about the behaviour of these largely unknown components. However, although this new current changed and spread the aspect(s) of interest in vocabulary research, it still treated vocabulary as knowledge of components individually. Consequently, the relationships among these word knowledge aspects is still unclear. A handful of studies attempted to investigate the interaction between components by concurrently measuring various vocabulary aspects. Despite their insightful information about L2 learners' general vocabulary knowledge and the connections between aspects, these studies are still initial attempts to the investigation of overall vocabulary knowledge, and hence are limited in the information they can offer. The inconsistencies in the factors and word knowledge components examined, as

well as the different ways these components were measured, mean that the findings of these studies differ and cannot be directly comparable. In addition, their research design, methodologies and the almost non-existent replications make it difficult to estimate the relationships and order of acquisition of these aspects relative to each other. Therefore, the understanding of overall vocabulary knowledge and how the various components relate to each other in the mental lexicon depends on the examination of multiple word knowledge aspects receptively and productively, by means of sophisticated statistical procedures that can better show their interactions and scrutinise the generalisability of the findings. The present thesis is devoted to this aim.

Chapter 3

Development of the measurement instruments for vocabulary knowledge components

The multidimensional nature of vocabulary knowledge involves understanding and operationalising it as more than just knowledge of the form–meaning link of a word. Although many scholars agree that vocabulary knowledge should be measured including a variety of tests of multiple word knowledge aspects, most research has measured these components individually. As a result, little is known about how they interrelate. In order to investigate this issue, researchers have claimed that a test battery designed to assess various components independently, but concurrently, is required (Henriksen, 1999; Li & Kirby, 2015; Milton & Fitzpatrick, 2014; Nation & Webb, 2011; Read, 2000; Schmitt, 1997).

The aim of this thesis is, precisely, to explore the knowledge that second language learners have of various word knowledge components in order to establish relationships between them. Due to the lack of standardised tests, I developed a test battery that measures multiple vocabulary components concurrently for a set of real words. The present chapter is devoted to explaining the development of this test battery. I start by presenting the requirements for the test design in terms of the target population and practical constraints. Then, the process for selecting the target words and vocabulary knowledge components to be tested is introduced. Finally, I explain the piloting analyses conducted to devise the final test battery.

3.1. Design considerations

The experimental design of this test expanded on earlier studies that have examined multiple word knowledge components concurrently (Chen & Truscott, 2010; Schmitt, 1998; Webb, 2007a). Yet, the different aims and wide scope of this thesis

required certain compromises and thoughtful consideration with regard to aspects such as the selection of target words and components to be assessed. These are explained in the following sections.

3.1.1. Target population

One of the first factors to take into account in the design of the test battery was the participants' sample and its characteristics. In this regard, an important feature to acknowledge was the number of participants needed. The statistical technique to be used in Studies 3 and 5 (i.e., Structural Equation Modelling) requires a minimum of 100 participants for the results to be reliable (Kline, 2016). Thus, the target was to collect data on the knowledge of vocabulary components from around 100+ participants in each study. This requirement involved taking into consideration the participants' linguistic background. Due to the foreseen difficulty of finding the required number of learners willing to complete a predictably long test, I first contemplated the possibility of assessing participants from different mother tongues to increase the chances of obtaining the number of participants needed. However, analysing data from learners of mixed first languages (L1s) would involve too many factors influencing the results in different ways, and thus making the findings difficult to interpret. Therefore, for the aim of this thesis, I considered it more appropriate to focus on a homogenous group of participants from only one mother tongue in each of the studies. This way, the test battery could be specifically designed to attend to the particular characteristics of the target population.

A final factor to consider was the participants' proficiency level. The method of analysis to be employed in Studies 1 and 4 (i.e., Implicational Scaling) requires participants to have different levels of vocabulary in English. The target population for each study, thus, included learners of English as L2 with a range of vocabulary proficiencies, from beginners to advanced. Consequently, the test battery had to be designed to accommodate to the different vocabulary knowledge proficiencies of participants.

3.1.2. Selection of target word knowledge aspects

The aim of the test battery is to measure knowledge of multiple vocabulary components. Thus, ideally all components specified in Nation's (2013) framework of what it means to know a word should be tested. However, this is an unrealistic target. Therefore, decisions had to be made regarding the number of aspects that could be assessed, and which ones would provide a better representation of general vocabulary knowledge.

One possibility to develop a test battery of reasonable duration was to test a larger number of word knowledge aspects but only at one level of sensitivity (i.e., either in recall or recognition). On the other hand, fewer aspects could be tested but both in the recognition and recall masteries. Many researchers have asserted that different levels of sensitivity need to be measured to achieve more precision in the assessment of depth of vocabulary knowledge (Cheng & Matthews, 2018; Henriksen, 1999; Nation, 2013; Nation & Webb, 2011; Webb, 2005). Therefore, the decision was made to compromise on the number of aspects to be tested, constraining it to four[4], and test them both in recognition and recall.

Another important factor to take into consideration was the mode of presentation of the components. Previous research has shown that spoken knowledge of vocabulary aspects generally lags behind the written knowledge in adult L2 learners (Milton & Hopkins, 2006; Milton & Riordan, 2006). Thus, measuring both modes across different components would provide a more accurate picture of overall vocabulary. Nevertheless, assessing recall and recognition knowledge of multiple components in both spoken and written mode would be impracticable. Therefore, for practical reasons I decided to measure the aspects only in the written mode, although I acknowledge the limitation of doing so. Yet, this decision allowed me to measure more vocabulary components, since spoken tests tend to take longer.

As shown in Table 1, Nation's (2013) framework divides vocabulary knowledge into 3 main constructs, *form*, *meaning* and *use*, which, in turn, involve other aspects within them. When selecting the specific vocabulary knowledge aspects to be tested in my studies, I considered it essential to have at least one representative of each of these

[4] Testing four aspects in recall and recognition was considered the maximum amount of possible aspects to be assessed for time constraints. A target of 20 words per aspect would allow 25 minutes on average to complete each task, plus 5 minutes per test to rest. This would lead to a 3.5-hour test plus 30 minutes for a break.

three bigger constructs. Following Nation's approach, the construct *form* includes knowledge of the spoken form, written form and word parts of a word. Although it is possible to have some knowledge of the form of a word without actually knowing its meaning, such as knowing that the word exists in a language, pragmatically the distinction between form and meaning is very difficult to operationalise. Thus testing form alone (whether spoken or written) independently of meaning was considered unfeasible and not very informative for the purpose of this thesis, since it is not enough knowledge for using a language. Conversely, having morphological knowledge of a word has been found important for communication and to help vocabulary growth (Kieffer & Lesaux, 2012b; Schmitt, 2010). Thus, I decided to test knowledge of word parts, in particular derivatives, to account for knowledge of the form of a word.

Regarding *meaning*, Nation (*ibid*.) suggests that in order to acquire full meaning-knowledge of a word learners need to know the form–meaning link, concepts and referents (i.e., its multiple meanings), and associations (i.e., paradigmatic associations such as synonyms or superordinates). Arguably, the most basic aspects to acquire comprehensive knowledge of the meaning of a word are the form–meaning link and knowledge of multiple meanings. Knowledge of the form–meaning link is, in my view, the lowest meaningful word knowledge level that enables learners to use a language. For its part, knowledge of multiple meaning is beneficial to employ a word accurately and appropriately in different situations. Besides, while associations have been more commonly researched (Fitzpatrick, 2006; Fitzpatrick & Izura, 2011; Meara, 2009), multiple-meanings knowledge is rarely addressed in research or the language classroom. For these reasons, as a representation of meaning I was interested in examining knowledge of the form–meaning link and of multiple meanings of words.

Finally, grammatical functions, collocations and constraints on use are aspects involved in knowledge of the *use* of words (Nation, *ibid*.). Although knowledge of the grammatical functions and use constraints of a word are important for communication, they tap too much into other areas of language, such as grammar and socio-cultural-pragmatics. Even though I am aware that vocabulary can never be completely isolated from other areas of language, from these three use aspects I considered collocations to be a better representative of vocabulary knowledge, and thus it was the aspect chosen to embody the *use* construct in my test battery.

As a result of these decisions, the vocabulary knowledge test battery described in this chapter assesses recall and recognition knowledge of Derivatives (as a representative of form), Multiple Meanings (as a representative of meaning), Form–meaning link (as the connection between form and meaning), and Collocations (as a representative of use).

3.1.3. Selection of target words

A particularly important matter in the design of the test battery involved the selection of the stimulus words. Due to the fact that multiple word knowledge components were examined, the target words required meeting specific criteria.

Firstly, for a pattern of difficulty of word knowledge aspects to be found, the implicational scaling procedure demands the linguistic items tested to have different difficulty levels. In order to account for these different levels of difficulty, the target words were selected from a range of frequency bands. A norming study with 30 sample-like participants indicated that their vocabulary level was in the 3,000-5,000 frequency bands on average. Thus, the first criterion for the target words was that they ranged between the 1,000 and 10,000 frequency bands. Yet, research has found that not all frequency levels are equally important, and that the frequency effect decreases at lower frequencies (Aizawa, 2006; Schmitt & Schmitt, 2014). For this reason, it has been suggested that vocabulary tests should use more fine-grained bands at the high-frequency end and sample more items from these crucial bands (Kremmel, 2016). Thus, although the target words in the test battery included a range of high-, mid- and low-frequency bands, more words were sampled from the higher-frequency levels. This way, not only the requirement of different difficulty levels of words was met, but I also ensured that lower-proficiency learners could manage the test battery.

Secondly, a central consideration in the test design comprised whether to assess the same set or different sets of target words across aspects. Assessing a list of different words in each vocabulary component would prevent any influence of having encountered the word in previous tests on subsequent ones. However, this decision would bring up other issues. First, Nation's (2013) framework of vocabulary knowledge specifies that every word needs to be learnt across the various aspects. If this is the case, we need to examine how a word develops along the different word knowledge

components. Selecting a different set of words for each aspect would not allow this examination. More importantly, this approach would affect the interpretation of results for each aspect. Testing different stimulus words would mean that the knowledge of each aspect would be influenced by the different particular characteristics of the tailored sets of words chosen in each test. Therefore, the knowledge of the different vocabulary components could not be comparable, and as a result the relationship between aspects would be altered. Since one of the main aims in this thesis is to examine how knowledge of the different components develops in order to establish the relationships between and order of acquisition of these components, this solution was inappropriate. Using a different set of words for each component would increase the risk of the scale of acquisition being an artefact of the target words selected for each aspect. Consequently, the most suitable solution for the purpose of this thesis was testing the same set of target words across all vocabulary components. This allowed me to investigate the extent to which different aspects have been acquired for each target word, and ensures that each component is being measured in a consistent and balanced manner in terms of the characteristics of the target words. Although this decision also comes with some consequences (i.e., potential contamination from one test to the following ones), it is the approach that most previous multicomponent studies have adopted, confirming its appropriateness for the purpose of investigating overall vocabulary knowledge (Chen & Truscott, 2010; Chui, 2006; Schmitt, 1998; Scott, Flinspach, Vevea, & Castaneda, 2017; Spencer et al., 2015; Webb, 2005, 2007a, 2009).

Finally, deciding how many stimulus words to assess was prompted by the limitations of previous research and the length of the current test. One of the major criticisms of the few studies assessing various word knowledge aspects is the small number of words used (e.g., Webb, 2005, tested ten words, while Schmitt, 1998, tested 11). Although the reason for the selection of fewer words is to compensate for the length of the total test when assessing various word knowledge aspects, it is necessary in depth of vocabulary studies to examine a more representative sample. Therefore, testing more than ten words was the goal in my test. Schmitt et al. (2001) suggest that 30 words is the minimum number for a good estimate of the vocabulary profile in a frequency level when assessing the form–meaning link. Nevertheless, measuring depth of knowledge involves that the test administration is time consuming, which limits the number of target words that can be tested. Assessing 30 words in four different word knowledge

aspects in recognition and recall would increase the length of the test to an unfeasible target, especially considering the large number of participants required (100+). Moreover, the purpose of the thesis meant that it was better to compromise on number of words than number of vocabulary components. Therefore, the final test battery tested 20 target words across multiple components at two levels of sensitivity.

3.1.4. Criteria for stimulus words

Following the test design considerations and requirements described above, the list of stimulus words was developed satisfying the following criteria:

- ***A range of frequencies***. Words from various frequency bands were included in the stimulus list. This permitted control of the items' different degrees of difficulty as well as of the participants' diverse proficiencies, as required by the implicational scaling analyses. The final list of target words comprised words from the 1K to 9K frequency bands, namely seven words were in the 1K, six in the 2K, two in the 3K, and one word in each of the remaining bands 4K, 5K, 6K, 8K and 9K (see Table 3).
- ***A range of parts of speech***. Words from different word classes had to be included in the test in order to have a representative sample and to control for the effect that selecting only one part of speech can have in their knowledge, as they could be either particularly easy or difficult for students (e.g., Dóczi & Kormos, 2016). The word class in which words were tested was selected on the basis of the most frequent sense of that word. For example, the polysemous target word *indent,* which can be both a verb and a noun, was counted as a verb in the part of speech ratio because, according to different reference dictionaries (see Footnote 5), its most common meaning is as a verb. Since I am testing derivative knowledge, the specific word class of the target words was used as the headword to assess the particular item in the other word knowledge aspects. The final stimulus list included 8 nouns, 7 verbs and 5 adjectives (see Table 3).
- ***Multiple-meaning words***. In order to assess knowledge of multiple meanings, each target word needed to have at least three senses as described by various reference dictionaries (see Footnote 5). The different meanings were targeted to be as different from each other as possible, although the multiple criteria needed

to be met by the target words meant that for some words the various meanings were closer to their core meaning than for others (e.g., *hard* as: 1) not easy to cut; 2) difficult to understand; 3) needing a lot of effort). Moreover, in this thesis, I am interested in words that have the same form and multiple meanings, disregarding the origin of the different senses. Therefore, the target words included both polysemes (words whose meanings derived from the same origin and are, therefore, more similar to each other) and homonyms (words that evolved on the same form but come from different origins, and thus have very dissimilar meanings).

- ***At least three derivative forms for one sense***. Since knowledge of derivatives is one of the targets of this thesis, all the stimulus words needed to have at least three derivative forms for the same meaning. An interesting finding during the selection of the target items was that words with multiple meanings usually have different derivative forms depending on the meaning. For example, the word *appropriate* as an adjective, meaning 'suitable', has the adverb *appropriately*, but no noun form, whilst *appropriate* as a verb, meaning 'to take or steal', has the noun *appropriation*, but not an adverb form. Hence, for one meaning a headword can have a set of derivatives somewhat different from those of another meaning, and different missing word class forms depending on the sense. For this reason, it was a requirement for the target words in the test battery to have at least three derivatives for the main meaning tested.
- ***A proportion of cognate and non-cognate words***. Cognate words are common in English, and have been found to influence the acquisition of words (De Groot & Van Hell, 2005; Laufer, 1997; Rosselli et al., 2012). Therefore, a percentage of Spanish-English cognates representative of the proportion in general English was included in the stimulus list. Some indications are that the percentage of Spanish-English cognates is around 34-37% (Lubliner & Hiebert, 2011; Wallace & Hammill, 2002). Therefore, the target words included 35% cognates (n = 7) and 65% non-cognate words (n = 13). Since the target samples in the studies were English learners from a cognate language (i.e., Spanish) and from a non-cognate language (i.e., Chinese), the results in the different tests can be compared as a function of cognateness.

Due to these strict criteria for target words, it was not possible to employ any available frequency list (e.g., 5,000 Corpus of Contemporary American English (COCA) list (Davies, 2008)) to select the final target items. Thus, I had to make a more informed selection of words. The most difficult criteria to satisfy were regarding the derivatives and multiple meanings, since all words have a form–meaning link and collocations. Yet, not all words in a language possess three distinct meanings or three derivatives for one meaning. Thus, I started the development of the target word list by finding multiple-meaning words with the required number of derivatives. Then these were classified according to their frequency level, part of speech and cognateness status. Finally, a frequent collocate of those words was selected. The target words were checked in three main Advanced Learner's dictionaries[5] to ensure all the criteria were met. After these procedures, a selection of 29 words were submitted to piloting by both native and non-native speakers (see Section 3.3.1 for details on the piloting procedure), from which 20 final target words made it to the test battery (see Table 3). These target words have ecological validity in that they represent the different types of items students may encounter (e.g., cognate words, words from different word classes, different frequencies, and different types of multiple meanings).

3.2. Measures

The measurement instrument used for data collection in this thesis consisted of a battery of pen-and-paper tests to retrieve information about the knowledge of various vocabulary components at two levels of sensitivity: recognition and recall. Following Read's (1993) characterisation of vocabulary test formats, the tests used in this battery required demonstration of knowledge by participants. Also, for the sake of ecological validity, the target words were tested in context rather than isolated. Therefore, in all the tests I employed short sentences to contextualise the words, trying to make sure that the amount of information in them was similar for all words and not too informative. Following Webb (2005), this contextual sentences were taken and/or adapted, when possible, from examples in the corpus-based dictionaries used as reference. Finally, in order not to hinder low level participants, the words used in these prompt contexts (as

[5] *Cambridge Online Advanced Learners Dictionary*, *Longman Dictionary of Contemporary English* (2003), and *Collins COBUILD Advanced Learner's English Dictionary* (2006).

well as in the instructions to the tests) were from the first 2,000 most common words (determined by *Vocabprofile* on the *Lextutor* website, Cobb, 2015).

After a thorough critique of the types of tests used in the literature to assess knowledge of the various word knowledge aspects, I selected the following test formats for each aspect measured (see Table 2).

Table 2 Test formats per word knowledge aspect.

Test	Format
Form–Meaning recall	Context plus filling in the gap in a sentence.
Form–Meaning recognition	Multiple-choice test (five options, one correct answer).
Derivative recall	Filling in the gap with the appropriate root plus suffix for each word class.
Derivative recognition	Multiple-choice test (eight choices, four correct answers).
Multiple-meanings recall	Open-ended description/definition of each meaning per word.
Multiple-meanings recognition	Multiple-choice test selecting the sentences in which the word was being used in the correct sense (six options, three correct answers).
Collocation recall	Context plus filling in the gap with the appropriate collocate in a sentence.
Collocation recognition	Multiple-choice test (five options, one correct answer).

Since the different word knowledge aspects required different test formats, each individual test started with their specific instructions and illustrated examples on how to complete them. Apart from the battery of tests, a questionnaire eliciting background information about the participants and a test to assess their general vocabulary level (VLT, Schmitt et al., 2001) were used (see Appendix 3 for the complete test battery and the background questionnaire).

The individual tests were designed so that they could not be answered based on the information in previous tests. For example, following numerous previous studies (Chen & Truscott, 2010; Chui, 2006; Schmitt, 1998; Webb, 2005, 2007a), the recall tests were always given before their recognition counterparts. Furthermore, in order to

minimise any memory effect, I ensured that the targeted derivatives, meanings and collocations did not appear in the test before the section in which they were measured, and that the contexts and sentences in which the words were used in the earlier tests were different than the ones in the later tests. Also, I inserted parts of the VLT between recall and recognition sections of some word knowledge components to minimise any possible cross-section contamination (see piloting Section 3.3.2 for details).

Finally, the different formats and characteristics of the individual tests required each of the four aspects to have a different scoring system. Partial knowledge for the different aspects was not addressed due to the difficulty of defining what partial knowledge is, and only full scores were given for each item. Following the instructions in the tests, when a question was left unanswered it was considered lack of knowledge and coded as 0. The next sections will describe the development of each test, as well as the scoring protocols employed for each of them.

3.2.1. Form recall knowledge of the form–meaning link

This test followed a fill-in-the-blank format, where participants were asked to recall the L2 form, given the L1 meaning (e.g., Laufer & Goldstein, 2004; Webb, 2005). This is considered the most difficult degree of knowledge of the form–meaning link (Laufer & Goldstein, 2004). The most frequent sense of each word was tested, as indicated by the three learner dictionaries used as reference. The participants were presented with a context in their mother tongue (i.e., Spanish or Chinese) setting the situation and meaning of the target word. The example below represents an item for the target word *season* in the Spanish test, and translates as "Summer is the best time of the year for me, because I like the heat a lot and being able to go to the beach". In the case of cognate words for the Spanish participants (e.g., *terminal* and 'terminal'), the Spanish context was paraphrased so that the cognate word form did not appear and influence the answer. For non-cognate words (e.g. *employ* and 'contratar') the equivalent word would be used in the context to prompt recall of the target word's form.

Participants had to fill in the blank in the English sentence. The first letter in the gap was supplied to constrain the possible answers to the target word. Items were scored dichotomously as correct (given a score of 1) or incorrect (given a score of 0). An answer was considered correct if participants provided the correctly-spelled form of the

word or a version which was recognizable (e.g., *redem* for *redeem*, *distintion* for *distinction*), in either the British or American spelling. Responses were considered incorrect if they were left blank, the word was not the target word, or if it was hardly recognizable as the target word (e.g. *drefht* for *draught*). The maximum total score of this test was 20 points (one for each of the 20 target words).

> *El verano es el mejor periodo del año para mí, porque me gusta mucho el calor y poder ir a la playa.*
>
> It is my favorite s_____.

3.2.2. Meaning recognition knowledge of the form–meaning link

This test utilised the format that is perhaps the most widely used in research assessing this type of knowledge: a meaning recognition multiple-choice test (e.g., Laufer & Goldstein, 2004; Webb, 2005). The target words were presented in a short, non-informative sentence, and participants were asked to select the correct meaning from four options with three distractors, as in the example below.

Multiple-choice tests have been claimed to tend to overestimate one's knowledge (Gyllstad et al., 2015; Stewart, 2014). However, they are still widely employed in research due to their ease of administration and the students' familiarity with the format. In order to keep the possible guessing effects to a minimum, I took four actions[6]. Firstly, participants were encouraged not to guess blindly, both in the written instructions and the oral explanations. Secondly, three distractors were included for each item, as having more options to choose from makes guessing more difficult (Stewart, 2012). Thirdly, an 'I don't know' option was included, as research suggests it can prevent guessing (Zhang, 2013), and participants were instructed to choose this option if they did not know the answer. Finally, students were asked to indicate in a 1-4 scale how sure they were of their answers (1 = not sure at all; 4 = totally sure), in order to encourage reflection of their degree of knowledge (see Appendix 3). As in the recall test, the scoring was dichotomous. Responses were scored as correct only if the appropriate meaning was selected. Conversely, an item was scored with a 0 if a wrong

[6] These actions to minimise guessing were followed in all the multiple-choice tests in the battery (i.e., the recognition tests).

answer was chosen, if the question was left blank, or when the 'I don't know' option was selected (this applies to all the following recognition tests). The maximum total score for this test was 20.

It is the best *season*.

a) Animal
b) Time
c) Appearance
d) Place
e) I don't know

3.2.3. Form recall knowledge of derivatives

This task took the format that has been used in previous research analysing productive knowledge of derivatives (e.g., Saigh, 2015; Schmitt & Zimmerman, 2002). In this task, each of the 20 target words was tested for their derivative forms for the four main word classes: noun, verb, adverb and adjective. Participants were asked to write down the derivative forms of the target word that were appropriate in four English sentences written to constrain word class. The sentence was the same for all the word classes, although slightly adapted to fit the specific derivatives tested. Participants were informed that the form of a word sometimes does not change for different word classes, and that some words may not exist in all the word classes. As in the other tests, examples were provided to illustrate the completion of the task. The scoring of each target word ranged from 0 to 4 depending on how many of the four derivatives were known (e.g., for the target word *hard*: noun for *hard* =1, verb for *hard* = 0, adjective for *hard* = 1, adverb for *hard* = 0; total score = 2/4).

Target word	Noun	Verb	Adjective	Adverb	**Total Score**
Hard	1	0	1	0	2

Responses for each word class were scored as correct (1) if the suffixes needed to create the derivatives were correctly spelled (with the exception of double letters, e.g. *freshnes* for *freshness*, *meaningfull* for *meaningful*), even if the root had minor spelling errors (e.g. *chalenging* for *challenging*, *fullfilling* for *fulfilling*). It was also scored correctly if the participants successfully indicated with an *X* that a word class did not

exist. An item was considered incorrect in three scenarios: 1) the suffix added was erroneous, 2) if an *X* was written where a word class form existed, or 3) if gaps were left blank, as specified in the instructions.

The sentences for the verb forms were written so that the infinitive of the verb was the correct option. If participants wrote the verb in the past in this section it was scored as incorrect, since the *–ed* suffix was often the correct form for the adjective word class, (e.g. *close* – adjective *closed*; *grate* – adjective *grated*). Finally, when participants provided some derivative forms that were not the ones targeted but were correct in the sentence according to native-speakers' intuitions and the COCA information, they were accepted as correct. For example, the adjective forms *developing* and *developmental* fit in the sentence, and then both were accepted as correct answers. Although the targeted answer was *developmental*, native raters considered *developing* as suitable in the sentence and was confirmed with COCA information; thus, it was accepted as a correct response. Nevertheless, *developing* showed to be much less common than *developmental*, according to COCA information. Since every word in derivative knowledge was scored out of four points (one per word class), the maximum score for this test was 80 (4 points x 20 target words).

Season

Noun	In this country, each _____ is clearly different.
Verb	In this country, the temperature variations _____ clearly.
Adjective	In this country, the _____ temperature variations are clearly different.
Adverb	In this country, the temperature variations occur _____.

3.2.4. Form recognition knowledge of derivatives

The design of this test was based on the format used by Saigh (2015), and adopts a multiple-choice task with multiple answers. As in the recall test, recognition knowledge of derivatives was assessed by testing four word classes per target word. The learners were presented with eight different derivative options for each target word. Generally, this would include one correct derivative and one distractor per word class. The distractors were carefully designed by developing a list of the most commonly-used

suffixes for each word class, and applying them to the root of the target words to create invalid forms of each word class. These invalid forms were piloted by asking native and non-native speakers which ones seemed plausible in English. Because some target words do not have derivatives for all four word classes, the option *X* was given in all items.

In some cases a word class would only have one derivative option. For example, the target word *Fresh* has forms for all the four word classes. Since I have to include always the option *X* (not to give away too much information about the existence or lack of existence of a word class for a word), I inserted two options for the noun (the correct one (*freshness*), and a distractor (*freshment*)), two for the verb (*freshen* and *freshate*), two for the adjective (*fresh* and *freshable*), and one for the adjective (*freshly*). For those target words that did not have derivatives for all the four word classes (generally being the adverb which was lacking) I still included a distractor form for the adverb and the option *X*, so that the items looked similar in all the target words. Participants were told to employ the *X* option when they thought that the word class for a particular target item did not exist, and to leave the item blank if they did not know the derivative form. A response was considered correct if the correct derivative form was selected (or *X* if appropriate) for each word class. If the right four derivative forms of a word were selected, but they were allocated to the wrong word class it was considered incorrect (e.g., selecting *fresh* as the noun, *freshly* as the adjective, *freshness* as the verb and *freshen* as adverb). The maximum score was 80.

Season

a. *Season*	***b.*** *Seasonize*	***c.*** *Seasonally*	***d.*** *Seasonation*
e. *Seasonate*	***f.*** *Seasonal*	***g.*** *Seasony*	***h.*** *X*

Noun	In this country, each _____ is clearly different.
Verb	In this country, the temperature variations _____ clearly.
Adjective	In this country, the _____ temperature variations are clearly different.
Adverb	In this country, the temperature variations occur _____.

3.2.5. Meaning recall knowledge of multiple meanings

A typical way of testing knowledge of multiple meanings in the literature is by means of an interview in which learners define the words' meanings orally (Schmitt, 1998; Tannenbaum et al., 2006). This method has the benefit of interacting with the student, being able to help them retrieve the knowledge if existent. Yet, the interview was an unfeasible format in my studies for three main practical reasons: 1) needing to collect data from 100+ students on multiple word knowledge aspects, a personal interview would have made the data collection impractical in terms of time and resources; 2) changing the format of the battery from written to oral and then written again (to follow the necessary order of administration of the test) would have been confusing for the participants and logistically unworkable; 3) all students would have needed to be assessed individually, making the data collection for such a number of participants impossible. Consequently, I designed a written test format in which students obtained enough information to be directed to the target senses of the words without providing too many details that enabled guessing. Moreover, learners were encouraged to ask the researcher (present during test administration) for reassurance if they thought they knew the meaning but were not sure it was the one being targeted.

This task involved a written open-question format, in which students were assessed on three meanings of each target word. Various meanings were piloted with five English native speakers to select the three meanings of a word that were the most discernible from each other. In this test, learners were given the target word, plus the word class, and a hint about each of the three meanings tested. After each clue, they had space to write, in their L1 or L2, a translation, a synonym, a description, a definition, or a sentence in which the specific meaning tested was used clearly. Answers were scored dichotomously, and participants received one point per sense correctly defined. Hence, each target word could have a maximum of three points. A response was scored as correct if participants provided a correct synonym, translation or description of that meaning, either in English or Spanish (e.g., producing *estación* [= *season*] as an answer to the first sense in the sample item below, writing 'adding salt and pepper to food' in the second sense, or writing 'en celo' [=in heat] in the third sense). Conversely, if participants left it blank or the description was too general and/or did not reflect the meaning of the word, the answer was considered incorrect (e.g., producing 'to cook typical food of a particular season' as an answer to the second sense in the sample item,

or ‘animales de verano’ [= summer animals] in the third sense). The maximum score for this test is 60 (3 points x 20 words).

Since this task involved scoring open questions, three raters were trained and asked to score a random 20% of the tests (see Chapter 4, Section 4.4.1.1, for the inter-rater reliability scores of the Spanish data, and Chapter 7, Section 7.4.1.1, for the scores of the Chinese data).

Season

(Noun= year)_____

(Verb= cooking)_____

(Noun= animals in season)_____

3.2.6. Meaning recognition knowledge of multiple meanings

This measure was adapted from a format used by Li and Kirby (2015). In this format, each target word was presented in five different sentences, with a different meaning in each. Three of those sentences represented the three meanings tested in the recall test, and in the other two sentences, the word was used with an invented meaning, acting as distractors. The distractor sentences were grammatical, with the word being used in the same word classes and the right sentence position. Irrelevant clues were avoided in the correct sentences, and both correct and distractor sentences were roughly of the same complexity and length.

The clue words from the recall test were not used in these sentences to minimise any memory effect. The exception was a single meaning sense from five words (*fresh* [water], *mean* [mathematics], *character* [writing], *bank* [money], and *grate* [teeth]) where it was difficult to write a clear, sensible sentence within the 2K word frequency band without the clue word. However, post-analysis showed that the mean recognition scores for these five meaning senses (2.15 out of 3) were virtually the same as the average of the other 55 recognition scores on the test (2.12), indicating that no cross-contamination from the recall test was in effect. The participants had to select in which sentences the target word was being used with a correct meaning, being told that there

was a minimum of one and a maximum of three correct sentences in each item[7]. An 'I don't know' option was given to avoid guessing, and was to be selected if the participant did not know any of the senses. One point was given per each correct sentence selected, with a maximum of three points per target word. The maximum score for this test is 60.

Season

a) The four seasons are winter, spring, summer and autumn.
b) The car´s season breaks very often.
c) Their dog is in season and can´t go out.
d) It is important to check the season of your computer once a year.
e) I forgot to season the fish with salt and pepper.
f) I don´t know

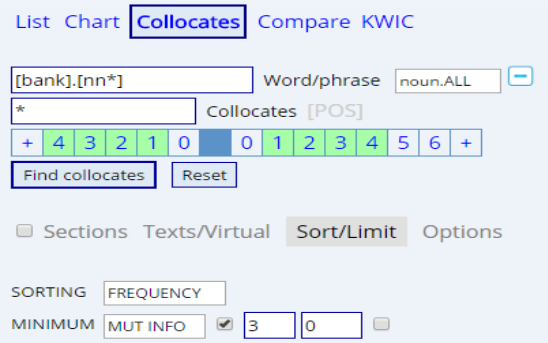
3.2.7. Form recall knowledge of collocations

Learners' form recall knowledge of collocations was assessed using a test format based on previous research (González Fernández & Schmitt, 2015; Peters, 2016). Participants were given a short context in their native language (i.e., Spanish or Chinese) and had to fill in the English sentence gap with the appropriate collocate given the first letter. The example below is taken from the Spanish test, and translated as 'Peak season is when most people go on holiday'. The sentences were written so that there were no direct translations which the participants could use to answer the items without knowing the collocates. For example, 'en temporada alta' from the example sentence below literally translates as 'in high season' rather than the English collocate 'in *peak* season'. The collocates were retrieved from the COCA (Davies, 2008), and were among the most frequent collocates of the target word. The collocations were all lexical and defined as any two words in combination with a greater corpus frequency than could occur by chance, irrespective of their level of compositionality and/or semantic transparency, that have an MI score ≥ 3 (Hunston, 2002), and within a span of ±4 words.

[7] The number of correct answers in the multiple-meaning recognition test was constant: three correct answers, corresponding to the three meanings tested in the recall test. Yet, students were told that there was always one minimum and three maximum correct answers, varying from item to item, in order to prevent them from blind guessing.

All collocates were from the 2K and 3K frequency bands (17 within the 2K level and three within the 3K level). The MI and frequency scores were estimated for the lemma, not the word family. An example of the parameters used for extracting the collocations for the study are presented in Figure 3.

Figure 3 Search parameters for collocate retrieval.



Answers were scored dichotomously as correct or incorrect. Correct answers required the participant to provide the correct collocate with correct spelling (except for minor misspellings, e.g. *desease* for *disease*, *prommise* for *promise*, *susteinable* for *sustainable*). Answers were considered incorrect when a wrong or non-collocate was produced (e.g., writing *sum* instead of *score* as collocate of *mean*, or *letter* for *line* as collocate of *indent*). The maximum score for this test is 20.

En temporada alta es cuando más gente se va de vacaciones.

When you plan to go on holidays you should bear in mind that hotels are always more expensive in p_____ *season*.

3.2.8. Form recognition knowledge of collocations

This task employed a multiple-choice format, based on previous research (e.g., Chui, 2006; Nguyen & Webb, 2017; Webb et al., 2013). Students were presented with a sentence in which the target word was underlined and they had to choose the appropriate collocate from four options, plus an 'I don't know' option. Participants were told that, although all the words were possible in the sentence, they had to select the option that sounded the most natural in English. The distractors were all words of the

same word class as the collocate, within the 1-3K frequency bands, and were non-collocates of the target word as established by COCA information (MI non-existent or ≤ 1). The possible scores for each target word were 0 or 1, and thus the maximum score for this test is 20.

There are always more tourists in _____ *season*.

a) Main
b) Peak
c) Big
d) Top
e) I don't know

3.2.9. 'I don't know' option

The test-takers were encouraged, both in the written instructions and the oral explanations, not to guess blindly. Nevertheless, an 'I don't know' option was included in all the recognition tests, as research suggests that it can prevent uninformed guessing (Zhang, 2013). For these tests, participants were explicitly instructed to select the 'I don't know' option when they lacked knowledge of the answer for a particular item. I assumed that if participants used this option in some tests/items, then they were not deliberately trying to guess, making it more unlikely that they would attempt guessing in other items. Thus, if they provided a wrong answer in other items it would be because they genuinely thought it was the correct option, not out of blind guessing. To get a clearer idea of their guessing behaviour, I checked whether participants were making use of this option. Overall, 68.7% participants used this option at some point in the form–meaning recognition test, 78.8% in the derivative test, 62.6% in the multiple-meanings test, and 57.6% in the collocation test. Recent studies suggest that the 'I don't know' option use may vary depending on the participants' proficiency levels, with more advanced learners making less partially-informed guesses (McDonald & Asaba, 2016), or their personality type, with low-confidence learners being less prone to guessing even with true partial knowledge (Huchinson, 2015). Nonetheless, my data suggests that, overall, most of my participants used this option at some points. I also found that the same participant that would use it for some items would not use it for others in which they selected the wrong answer, even in the same test. I interpret it as the participants using the 'I don't know' option only when they totally lacked knowledge of the answer,

and providing a response whenever they had intuitions or partial knowledge about the answer, even if the intuitions were eventually wrong. I believe that answering due to partial knowledge and perceptions is clearly different from blind guessing (see McDonald & Asaba, 2016), and thus, I concluded that most of the participants in this thesis generally employed the 'I don't know' option as advised, and did not blindly guess in the recognition, multiple-choice tests.

3.2.10. Two scoring perspectives

The tests for each of the four word knowledge aspects had different maximum scores. The form–meaning and collocation tests were scored out of a maximum of 20 (that is, a point for each of the 20 target words). However, for the derivatives and multiple-meanings tests, participants had to provide more than one answer per target word. In the case of derivative knowledge, each of the 20 target words required four answers (that means, the general maximum score for the derivative tests would be 80 points). In the multiple-meanings test, each target word required three answers, which makes a total possible maximum score of 60 points.

I considered a different scoring possibility in which the information about the individual words would be taken into account. This scoring perspective involved constraining all tests to a maximum of 20 points (one per target word). In this occasion, the scoring for the form–meaning and collocation tests remained the same, but changed for derivatives and multiple meanings. In the case of derivatives, each of the 20 target words would be considered known if at least 3 out of 4 of its derivatives were answered correctly. As for multiple meanings, participants would be considered to know the multiple-meanings aspect of a word if at least 2 out of 3 meanings of each word were answered correctly. Although this scoring perspective may seem preferable to get detailed information about the individual-word vocabulary knowledge, it conflicted with the implicational scaling analysis requirements in Chapters 4 and 7 (see Section 4.3.2). In short, to consider an aspect known this analysis requires a cut-off point that is generally set at the 80% accuracy level of the total possible answers for a component. Therefore, employing the individual word scoring perspective in derivatives (3 out of 4) and multiple-meanings (2 out of 3) knowledge would mean applying an extra constraint to these tests. That is, for the former test, it would require learners to know

75% (3/4) of the derivatives per word first, and then from those the 80% of total answers in order to consider a component mastered. For the latter test, it would involve knowing 66.6% meanings per word (2/3), and from those the 80%. Thus, this perspective was considered to treat the tests for each word knowledge component differently. Since I wanted to make sure that the criterion for percentage of accuracy in the implicational scaling analysis was the same for all word knowledge components, that is 80% of the total possible scores, the first scoring perspective was applied.

Therefore, although this scoring system might miss information about the individual target words, it was considered to be the most systematic and accurate scoring protocol for the purpose of Studies 1 and 4: discovering an implicational scale of word knowledge components.

3.2.11. VLT and language background

Apart from knowledge of the vocabulary components, additional information about the participants was collected. Firstly, a background questionnaire was added at the end of the test battery (see Appendix 3 for the English version). It included closed and open-ended questions retrieving information on the students' age, gender, exposure to English, other languages known, and self-reported proficiency in English. Yet, proficiency self-ratings can be problematic, as they are affected by multiple psychological and cultural factors. Therefore, along with the self-reported language proficiency ratings, a more objective measure of the participants' general proficiency was also collected. Since vocabulary size seems to correlate with general language proficiency (Beck et al., 2013; Milton, 2013; Stæhr, 2008), the commonly employed Vocabulary Levels Test (Schmitt et al., 2001) was used to retrieve a general proficiency measure. Although the original intended use of this test was that of diagnosis of the students' ability to recognise the form–meaning link (Kremmel & Schmitt, 2017), it has been commonly used in research to provide an estimate of the vocabulary size of L2 learners (Peters, 2016; Vilkaitė & Schmitt, 2017). Four sections of the VLT were deemed interesting for the purpose of this thesis (the 2,000, 3,000, 5,000, and 10,000 word frequency bands). The results from the VLT test were employed for the statistical analyses, while the proficiency self-report data was simply used for participants' description.

3.3. Piloting

A number of pilot analyses were conducted with three purposes in mind:

1. To select the final list of 20 target words that would be included in the test battery.
2. To ascertain the best order of administration of the different sections.
3. To determine the duration of the tests.

3.3.1. Items piloting

Following the criteria specified in Section 3.1.4, a pool of 29 potential target words were selected for the piloting stage. From these, only 20 stimulus words would be included in the final test. An initial test battery comprising the 29 potential target words was created, and submitted to piloting by native speakers (NS) of English and by Spanish and Chinese speakers similar to the samples of the studies.

Firstly, five native English speakers completed the test battery for the 29 words to ensure the clarity and feasibility of all of test items and formats. They were asked to report about the naturalness of the sentences and items, the clarity of the instructions, and the appropriacy of the answers. In general, these participants considered the question items and formats clear and natural. Some adjustments in terms of rewording and clarification were performed. Whenever an item caused confusion or inconsistencies in the NS's responses, it was removed or reworded according to their suggestions. For example, the potential target word *deep* was considered to be problematic in the form–meaning receptive test, where some NS could not select the correct option from the suggested distractors (*big, wide, low, long*). Also, the word *accept* caused problems in the multiple-meanings test, where some NS could not see the different meanings for the word. Similarly, some NS could not identify the adjective and adverb forms for the word *diffuse* in the derivatives test. Consequently, these items were removed from the final pool.

Secondly, the test battery improved by the NS's comments was piloted with a group of non-native speakers, learners of L2 English. Five Spanish speakers (two living in the UK and three in Spain), and six Chinese EFL learners (four living in the UK and two in China) completed the test battery to account for test difficulty. Words that were

found too easy or consistently problematic were deleted from the pool. For example, the words *add* and *desolate* were removed from the final list because the participants had problems discriminating their different meanings.

The final 20 target words included in the test battery were those that were clear for the L2 English learners and in which English NS provided consistent and completely correct responses. See Table 3 for the final list of target words.

Table 3 List of target words (ordered by frequency).

	Target words	Frequency band	Raw lemma frequency in COCA	Cognate	Word Length
1	*Mean (v)*	1K	336,890	✗	4
2	*Close (v)*	1K	170,039	✗	5
3	*Hard (a)*	1K	158,413	✗	4
4	*Development (n)*	2K	106,307	✗	11
5	*Season (n)*	2K	98,979	✗	6
6	*Bank (n)*	1K	77,116	✓	4
7	*Challenge (n)*	2K	72,443	✗	9
8	*Character (n)*	2K	56,925	✓	9
9	*Fresh (a)*	1K	43,551	✓	5
10	*Bright (a)*	1K	34,400	✗	6
11	*Broad (a)*	2K	32,282	✗	5
12	*Employ (v)*	1K	21,780	✓	6
13	*Distinction (n)*	3K	12,500	✓	11
14	*Charm (n)*	2K	7,653	✗	5
15	*Terminal (a)*	4K	6,459	✓	8
16	*Fulfil (v)*	3K	4,251	✗	6
17	*Grate (v)*	5K	3,817	✗	6
18	*Redeem (v)*	6K	1,911	✓	6
19	*Draught (n)*	9K	226	✗	7
20	*Indent (v)*	8K	136	✗	6

Note: *v* = verb, *a* = adjective, *n* = noun.

3.3.2. Order of tests administration

The different sections (i.e., each individual test) in the test battery were designed to measure various aspects of knowledge of the same 20 target words. Controlling for the order in which participants receive the individual sections would help minimise the potential effect of previous sections on the subsequent ones. Thus, the order of administration of the different tests was piloted with the same Spanish learners described above to find which order was less prone to influence from previous tests.

For each aspect, the recall tests were administered before their recognition counterparts, in order to prevent tests being answered based on information from previous ones. Following this approach, two different administration orders of the whole test battery were piloted (see Table 4).

Table 4 Potential test-administration orders.

	Test order 1		Test order 2
1.	Form–Meaning Recall	1.	Form–Meaning Recall
2.	Form–Meaning Recognition	2.	Form–Meaning Recognition
3.	Derivative Recall	3.	Derivative Recall
4.	Derivative Recognition	4.	Derivative Recognition
5.	Multiple-Meanings Recall	5.	Multiple-Meanings Recall
6.	Multiple-Meanings Recognition	6.	Collocation Recall
7.	Collocation Recall	7.	Multiple-Meanings Recognition
8.	Collocation Recognition	8.	Collocation Recognition

The two test orders were identical in the first five sections, but differed in the last three, where the Multiple-Meanings and Collocation recall and recognition tests were alternated. Half of the Spanish and Chinese piloting participants completed the battery in the first administration order, and the other half following the second order. They were interviewed during and after the test on the influence that previous sections had on subsequent ones. Participants commented that, although they could remember

that the target words had occurred before, they could not recall information that helped them in subsequent tests. Comparison of the results suggested that the second order option distracted participants' attention from previous tests better than the first option, making it less susceptible to cross-test effect. Thus, the final test battery followed the second test order.

Moreover, to further minimise any possible memory of the target words from earlier sections to later ones, I inserted the different parts of the VLT test as distractors between the recall and recognition sections of some word knowledge components. Following participants' comments, two locations were found strategic to divert participants' attention from the earlier tests: between the form–meaning recall and recognition tests, and between the collocation recall and multiple-meanings recognition sections. Therefore, the order of administration of the different sections including the VLT breaking up the components test was: Form–Meaning link form recall → VLT 5K/3K → Form–Meaning link meaning recognition → Derivatives form recall → Derivatives form recognition → Multiple-meanings recall → Collocate form recall → VLT 10K/2K → Multiple-meanings recognition → Collocate form recognition. The 2K VLT level was placed at the end of the battery to put the easiest VLT section towards the end of a long test.

3.3.3. Test duration

Testing various word knowledge aspects at two levels of sensitivity (recognition and recall) implies that the test battery would take a very long time to complete. In order to get an estimation of the total length of the test, I piloted it with the same L2 English learners described above. The participants living in the UK took around 2.5 hours to complete the test, including some breaks. This suggests that high-proficiency learners could complete the test in that amount of time. The participants living in their home countries took one hour longer, completing the test in 3.5 hours. Thus, I estimated the duration of the whole test battery in about 3.5 to 4 hours altogether with some breaks. This would allow lower-proficiency learners enough time to complete the test.

Due to the duration of the test battery, finding volunteers to complete the test was envisaged to be complicated. Thus, the piloting participants were interviewed after the test about their preferred modalities of test administration, to attract more

volunteers. It was suggested that I gave an option to complete the test in two separate sessions. Thus, during the test battery administration, participants were given the option to take the test in one four-hour session or two non-consecutive two-hour sessions. For the latter scenario, the test was divided into two parts, with a break in between. Each part contained ten of the twenty target words, and were tested across all components. The two parts were designed so that they were equivalent, being balanced for frequency and cognateness status: both parts comprised words with a range of frequency levels and the same amount of cognates (see Table 5). The parts were counter-balanced, with some participants randomly assigned to do one set of ten words first, and other participants the other set first. In this case, the participants took the 5K and 3K VLT inserts in the first half of the test battery, and the 10K and 2K in the second half, but in the same positions described above. Table 5 shows the word division. For comparability, the test format designed to be completed in one session also had a counterbalanced presentation of words with two versions. In Version 1, participants completed a set of ten words in the first place, and in Version 2 the other set of ten words occurred first.

Table 5 Target words in each test section.

Version 1			**Version 2**	
Section 1	Section 2		Section 1	Section 2
Fresh (1K)	Bank (1K)		Bank (1K)	Fresh (1K)
Redeem (6K)	Grate (5K)		Grate (5K)	Redeem (6K)
Employ (1K)	Bright (1K)		Bright (1K)	Employ (1K)
Fulfil (3K)	Terminal (4K)		Terminal (4K)	Fulfil (3K)
Charm (2K)	Season (2K)		Season (2K)	Charm (2K)
Indent (8K)	Broad (2K)	⟷	Broad (2K)	Indent (8K)
Mean (1K)	Draught (9K)		Draught (9K)	Mean (1K)
Character (2K)	Challenge (2K)		Challenge (2K)	Character (2K)
Hard (1K)	Distinction (3K)		Distinction (3K)	Hard (1K)
Development (2K)	Close (1K)		Close (1K)	Development (2K)
4 cognates	3 cognates		3 cognates	4 cognates

Finally, since my sample required learners with different L2 proficiencies, participants were given the option to take the tests with the instructions in their mother tongue or in English. This would ensure that low-proficiency students were not hindered

by their limited English knowledge. Thus, the two test versions were developed with the instructions in English and in Spanish or Chinese[8]. This leaves a counterbalanced design with four versions in total for each study: Version 1 English, Version 2 English, Version 1 Spanish/Chinese and Version 2 Spanish/Chinese.

3.4. Cross-test influence

Measuring knowledge of various word knowledge components across the same set of words is the standard and best way of exploring the development of depth of knowledge of individual words, as attested by the previous multicomponent studies (Chen & Truscott, 2010; Chui, 2006; Schmitt, 1998; Webb, 2005). Yet, this approach implies showing the specific set of words in multiple contexts and situations, and thus some cross-test influence may occur. In order to minimise cross-section contamination, I took multiple steps. First, the different types of word knowledge in this study were conceptualised and operationalised so that they were isolated from each other as much as possible, tapping into related, but distinct, components. Also, the individual tests were designed so that the information in earlier tests could not inform the responses to subsequent tests. Moreover, I ensured that the targeted derivatives, meanings and collocations did not appear in the test before the section in which they were measured, and that the contexts in which the target words appeared were different in each test. Furthermore, the order of administration of each section was controlled to reduce the memory effect, both in terms of recall/recognition presentation and the specific components. Finally, the VLT parts were inserted between the different sections to divert participants' attention from test to test.

Although some degree of testing effect is probably unavoidable, I believe the various actions taken to prevent test effect suggest that the influence of previous tests on later ones is minimal. In order to provide evidence for this claim, I computed multiple analyses to evaluate the potential cross-test effect and validate the implicational scale established in Chapter 4 (see Section 4.5.3). This was done by means of McNemar's tests, which retrieved the number of people who mastered a later aspect without having mastered the previous ones. In order to check this, the aspects were divided into recall

[8] For space constraints, only the English version of the test battery is provided in this thesis (see Appendix 3). The Spanish and Chinese versions are available upon request.

and recognition, following the implicational scale. This influence was tested at the two cut-off mastery criteria employed in the implicational scaling analysis in Chapters 4 and 7 (see Section 4.3.2). Table 6 reflects the data based on the mastery criterion of 80% correct answers, while Table 7 shows this data for the 75% accuracy criterion. The data showed that very few participants showed mastery of one word knowledge component in the implicational scale without having mastered the previous one, confirming that the influence from measuring the same target words across multiple aspects was minimal. This provides evidence that the efforts to control for the cross-test contamination were successful, and thus the results were generally not affected by this effect.

Table 6 Proportion of participants mastering a later aspect but not a prior one (80% accuracy).

Tests in IS order	**% of Participants**
Form–Meaning Recog = 0; Collocation Recog = 1	6.25
Collocation Recog = 0; Multiple-Meanings Recog = 1	0
Multiple-Meanings Recog = 0; Derivative Recog = 1	3.47
Derivative Recog = 0; Collocation Recall = 1	8.33
Collocation Recall = 0; Form–Meaning Recall = 1	2.08
Form–Meaning Recall = 0; Derivative Recall = 1	4.17
Derivative Recall = 0; Multiple-Meanings Recall = 1	0

Note: *Recog* = Recognition; 0 = not mastered, 1 = mastered.

Table 7 Proportion of participants mastering a later aspect but not a prior one (75% accuracy).

Tests in IS order	% of Participants
Form–Meaning Recog = 0; Collocation Recog = 1	4.17
Collocation Recog = 0; Multiple-Meanings Recog = 1	0
Multiple-Meanings Recog = 0; Derivative Recog = 1	2.78
Derivative Recog = 0; Collocation Recall = 1	4.86
Collocation Recall = 0; Form–Meaning Recall = 1	4.17
Form–Meaning Recall = 0; Derivative Recall = 1	5.56
Derivative Productive = 0; Multiple-Meanings Productive = 1	0

Note: *Recog* = Recognition; 0 = not mastered, 1 = mastered.

3.5. Conclusion

Vocabulary knowledge is a very complex construct that involves the acquisition of many word knowledge aspects, generally known as depth of vocabulary knowledge. Measuring depth of word knowledge is difficult and time consuming, and thus, it has rarely been examined comprehensively. As a result, research needs to be done that investigates the knowledge that L2 learners have of multiple aspects of depth of knowledge concurrently. Since to date no consistent test exists that can assess overall vocabulary knowledge, I designed and developed a battery of tests to address this gap in the literature. This test battery measured recognition and recall mastery of twenty target words across four word knowledge components: form–meaning link, derivatives, multiple meanings and collocations. The piloting of the final measurement instrument suggested it was valid for use in the following studies.

Chapter 4

Knowledge and order of acquisition of vocabulary knowledge components by EFL learners: An Implicational Scale

The great complexity of vocabulary knowledge has led to a general lack of investigation of vocabulary depth. This limited research on the multiple components involved in knowing vocabulary translates into the uncertainty of how these word knowledge components behave and relate to each other in the mental lexicon of L2 learners. Therefore, there is a need to investigate the knowledge that learners of L2 English have of various word knowledge aspects in order to establish connections between them. In particular, it is important to examine whether, as it occurs with other aspects of language (i.e., grammar), the vocabulary knowledge components follow an order of acquisition. The study presented in this chapter will explore this issue, investigating multiple vocabulary components concurrently by means of a test battery in order to explore whether it is possible to describe a pattern of difficulty of these aspects. In other words, whether certain aspects of word knowledge are typically easier than others, and thus, being acquired in a consistent order.

4.1. Background of the study

To date, the sequencing of vocabulary teaching materials has been based on intuitions and observations of what is easy or difficult. Finding an order of acquisition of the different vocabulary knowledge components would provide a sound basis on which to organise these materials. For example, if it is found that one aspect of vocabulary learning is typically easier than another, this aspect would be introduced before more complex ones. Therefore, a consistent order of acquisition of vocabulary

components could have a direct effect in language teaching, syllabus design and material development.

However, the terms 'difficulty' and 'ease' of vocabulary acquisition actually refer to notions of a complex nature. In general, it alludes to the amount of effort required to learn vocabulary (Webb & Nation, 2017). As seen in Chapter 2, Section 2.2.4, there are numerous factors that influence this difficulty or ease that learners face in learning L2 words (Laufer, 1997). In this thesis, these terms are employed for the purpose of evaluating the learning challenge of the vocabulary components relative to each other, and thus, they are used to refer to the extent of the knowledge of the various word knowledge components that learners have. Accordingly, the higher the knowledge of one aspect across participants, the 'easier' this aspect will be considered to be, and vice versa. Based on this perceived/observed 'difficulty' and 'ease', easier components would be expected to be acquired earlier than more complex ones.

The statistical method required to establish a sequence of the acquisition of different word knowledge aspects considering their relative difficulty is called Implicational Scaling analysis. It is a technique that can help make predictions about how different linguistic features, and thus also the various aspects of word knowledge, are acquired. For this reason, it has been suggested to be a proxy for systematicity in language (Rickford, 2002). In this study, Implicational Scaling will be applied to ascertain whether a systematic pattern of acquisition of the different vocabulary knowledge components exists.

4.1.1. What is Implicational Scaling?

Implicational Scaling is the term employed in linguistics to refer to the scaling technique generally known in statistics as *Guttman Scaling* or *Scalogram analysis*. This technique was invented in 1944 by Guttman for the measurement of social attitudes, and is based on the assumption that the items being studied differ in the degree of difficulty they present to participants, and thus, they can be organised in a scale according to their difficulty level. Implicational scaling allows the establishment of systematic relationships between variable features, as well as the reconstruction of patterns of acquisition of these features from variable data (Richards, 1980). Therefore, this method of data analysis can determine whether specific linguistic items can be

ordered according to their difficulty level in a reliable manner, so that if a more difficult aspect is known, it implies knowledge of the easier aspects.

The main purpose of implicational scaling is to ascertain a one-dimensional continuum (i.e., scale) to measure a specific concept. That is, the idea behind it is that if we have a set of items that are measuring the same underlying construct, then they are interrelated and can be arranged in order of difficulty on a continuum. If the items measured indeed form an Implicational Scale (henceforth IS), this means that there are strong implicational or hierarchical relationships between them, so that *X* implies *Y* but not the reverse (Johnson, 2013). As stated by Gibson (1998, p. 36) "[i]mplication is the relationship between two variables such that we can make a generalisation about the absence or presence of one from the absence or presence of another". This way when a learner knows a specific item in the continuum we can predict they will also know the previous (easier) items.

Therefore, implicational scaling allows us to examine whether learners acquire particular linguistic items in a specific order. In order to apply implicational scaling, we need to create an "implicational matrix" or scalogram (Gibson, 1998), where one axis represents the participants involved and the other axis the linguistic items tested. In this matrix, the linguistic variables are ranked along the horizontal axis according to the order of difficulty, and participants are arranged relative to each other along the vertical axis in terms of the number of items they answered correctly based on a specific criterion. This matrix allows for the study of linguistic items in terms of their difficulty order and of participants in terms of their performance according to the scalability or goodness-of-fit of the scale (Johnson, 2013). Using the terms of Newman (2010, p. 211), the horizontal axis represents implicational relationships between variables (i.e., *intervariable scalability*) and the vertical axis implicational relationships among speakers (i.e., *interspeaker scalability*). The former is the focus of this study.

Given that implicational scaling is a method to organise categorical data, generally binary (Johnson, 2013), the variables of interest need to be coded dichotomously (i.e., Yes/No, 1-0) in order to create the implicational matrix or scalogram. In linguistics, the items being measured are usually coded as known/acquired (1 or Yes) or not (0 or No). Below is an example that illustrates this procedure.

In phonology studies there is an interest in finding out whether a specific language group is developing a variety of the standard language, through the study of use of different phonemes and the order in which they are used. If we were interested in the appearance of these different phonemes by different speakers, we can create a matrix with their results, 1 meaning the phoneme is produced, 0 meaning it is not produced (see Figure 4).

Figure 4 Example of implicational matrix.

Speaker	Ph1	Ph2	Ph3	Ph4	Ph5
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	0
4	1	1	0	0	0
5	1	0	0	0	0

In the scalogram in Figure 4 we see the two axes. The horizontal axis shows that the phonemes tested are arranged according to their difficulty, with the assumption that the larger the number of participants that show knowledge of one item, the easier its acquisition must be. The vertical axis, shows the speakers arranged according to their performance, so that the knowledge of more aspects a participant has, the higher in the rank it will be. From this matrix we see that the easiest phoneme is Ph1, and thus it would be expected to appear first in the variety of language under study. Moreover, we observe that speakers behave differently in their production of the specific phonemes, and that, according to this matrix, one could generalise that if a student produces one of the most complex phonemes (e.g., Ph4) then they would also produce all the easier ones (e.g., Ph1, Ph2 and Ph3).

This scalogram shows an ideal scenario, in which the order of items and participants fall into the perfect pattern where we can draw a line through the scalogram clearly dividing known (1) and unknown (0) items. However, in linguistic data, this perfect situation rarely occurs. There would always be exceptions where participants seem to know a difficult item but not some of the easier ones. These exceptions are called random errors, and are inevitable in social sciences. To account for this situation, Guttman (1944) proposed a series of model-fit statistics that allow for certain amount

of random errors (i.e., up to 10%) where we can still consider a scale valid (see Guttman Scaling Procedure, Section 4.3.1).

Following the assumptions of implicational scaling, if it is found that the data forms an IS, it means that the items belong to a one-dimensional continuum, can be ordered according to their difficulty and thus are expected to be acquired in that order.

4.1.2. Implicational Scaling analysis in linguistics

Implicational Scaling was introduced to linguistics by David DeCamp (1971) to study the phonological, lexical and morpho-syntactic variation of Jamaican Creole. It constitutes a useful tool for the study of languages, since the implicational patterns of linguistic features found can be used to create rules about how those linguistic features are acquired by a particular community (DeCamp, 1971; Hatch & Lazaraton, 1991). Implicational scaling has been considered particularly useful and informative to answer questions concerning linguistic variability (e.g., language variation and change) and systematicity in second language acquisition (specifically phonology and morpheme studies). Thus, these are the areas that have commonly employed this technique (Andersen, 1978; Dulay & Burt, 1974; Gibson, 1998; Johnson, 2013; Rickford, 2002).

ISs are generally connected to the idea of dynamic processes, that is, changes that are completed or ongoing. Thus, implicational patterns describe the process of a change at a specific point in time which some learners have reached, but others have not (Bailey, 1973). Bailey (1973, p. 82) reiterates: “what is quantitatively less is slower and later; what is more is earlier and faster”. That is, linguistic aspects for which participants show less knowledge are acquired slower and later, whereas those that show more knowledge by participants are easier and appear faster.

Although the use of implicational scaling was of central importance in the 70s, it has decreased in recent years. However, as Rickford (2002) states, this neglect of implicational scaling does not have to do with any shortcoming of the method, and thus many researchers continue to support its use now and for the future (e.g., Hatch & Lazaraton, 1991; Johnson, 2013; Newman, 2010; Rickford, 2002). The following section presents some applications of implicational scaling in SLA.

4.1.3. Implicational Scaling in SLA research

Implicational scaling was first applied in linguistics for the analysis of syntactic variation (e.g., Andersen, 1978; Rickford, 1975) and phonological errors (Dickerson, 1975; Gatbonton, 1977). However, the field that has benefited the most from the use of implicational scaling analysis in the last 40 years is SLA, specifically the areas of morphology and phonology.

Regarding the acquisition of grammatical and morphological structures, there have been multiple implicational scaling studies indicating and replicating, in various situations and conditions, what seems a robust acquisition sequence of grammar rules and morphemes (Andersen, 1978; Bailey, Madden, & Krashen, 1974; Dulay & Burt, 1974; Ellis, 1989; Larsen-Freeman, 1975; Pienemann, 1998; Politzer, 1976). For example, Politzer (1976) used implicational scaling to explain the acquisition of five grammatical rules that differ in French and English. Andersen (1978) employed implicational scaling to study and describe the order in which Spanish-speaking students acquired 13 morphemes in English. Ellis (1989) studied the acquisition of German word order rules by adult language learners in a classroom setting to compare the sequence of acquisition of these rules with the sequence followed by learners in a naturalistic setting. He used implicational scaling to retrieve this order, and found no difference in the acquisition sequence in the two different settings. Pienemann (1998) applied the scaling model to investigate the mastery of morphological rules in German by non-native speakers, as a way to validate the Processability Theory (a psycholinguistic theory of L2 language development which proposes a particular sequence of stages learners go through when learning to produce syntactic structures) (see Chapter 2, Section 2.1.2.4 for more details). Ellis (2008) explored the difficulty of certain grammatical aspects by L2 learners, employing implicational scaling to test whether explicit knowledge of four grammatical structures follow the same difficulty pattern suggested by Pienemann's (1998, 2005) Processability Theory of implicit knowledge. He found that his implicit knowledge data followed Pienemann's predicted order, but this order was not supported by the explicit knowledge data.

These applications have inspired the use of implicational scaling methodology in other aspects of SLA. Nevertheless, there are only a handful of studies that have applied this technique to the study of vocabulary acquisition (see Gitsaki, 1996; Laufer & Goldstein, 2004; Mizrahi, 2016; Schmitt, 1997). The last two included different

aspects of vocabulary acquisition, whereas the first two studied only one aspect (collocational development and the form–meaning link, respectively).

Gitsaki (1996) studied the development of English collocational knowledge by Greek learners of ESL at three different proficiency levels. She found that there were stable patterns of development in the acquisition of collocations across and within the different proficiency levels. She concludes that the general development of collocational knowledge is as follows: first, students know simple and frequent grammatical collocations but use few types and repeat some of them a lot. At this stage, lexical collocations are known more accurately than complex grammatical collocations, but their accuracy is very low. Later, students start using more collocation types, using both simple and complex grammatical collocations although still at a very low accuracy level. Finally, at the higher proficiency level, students improve their accuracy with respect to both grammatical (simple and complex) and lexical collocations, acquiring an advanced collocational knowledge. In sum, this study shows how collocational knowledge seems to develop in L2 learners of English.

Laufer and Goldstein (2004) claim that the acquisition of the form–meaning link aspect requires mastering four degrees of strength, which they label as passive recognition (i.e., meaning recognition), passive recall (i.e., meaning production), active recognition (i.e., form recognition) and active recall (i.e., form production). After hypothesising an order of difficulty of these four degrees of strength of knowledge, they applied implicational scaling to the knowledge that 119 L2 learners of English of various L1s had of 16 words from different frequency levels. They found an IS which suggests that the order of difficulty of the form–meaning link levels is (from easier to more difficult) meaning recognition > form recognition > meaning recall > form recall.

Mizrahi (2016) investigated whether different measures of productive vocabulary knowledge and use constitute a scale of attainability to a nativelike level by highly advanced L2 learners and NS of English. The lexical variables under study were lexical richness, lexical variation, collocation use, vocabulary size, and collocation knowledge. She found that these indeed formed a reliable IS, so that it is easier to achieve a nativelike level in lexical richness, followed by variation and collocation use, with productive vocabulary size and collocation knowledge being the most difficult areas to attain nativelikeness (although the difference between these two was minimal).

Another study showing the application of implicational scaling in general vocabulary is Schmitt (1997). He looked at how different types of word knowledge (i.e., associations, spelling, grammatical knowledge, collocations and multiple meanings), were acquired. Because his study was the first attempt to submit different word knowledge components to implicational analysis, he employed various criteria that differed in strictness to develop the scales. Unfortunately, no IS was found in any of the criteria used. These results do not necessarily mean, however, that the various word knowledge components do not follow an implicational pattern of acquisition. As he suggests, the use of more precise measures of the word knowledge components and the testing of a larger number of participants may lead to statistically robust IS in the acquisition of word knowledge components. Following this, the present study aims to examine whether a pattern of acquisition of word knowledge components can be found.

4.2. Purpose and research questions

This study will investigate the knowledge that Spanish L2 learners of English possess of various vocabulary components, specifically form–meaning link, derivatives, multiple meanings and collocations. These types of knowledge will be assessed concurrently by means of a specifically designed battery of tests (see Chapter 3). The main objective of this study is to find out whether there is an order of acquisition of the different word knowledge aspects. The following research questions will be addressed:

1. How well are the different dimensions of word knowledge known by Spanish learners of L2 English?
2. To what degree are these dimensions interrelated?
3. Are some components of vocabulary knowledge typically acquired before other components? If so, is it possible to outline an order of acquisition of these components?
4. If an order of acquisition is found for the different aspects of word knowledge, does it remain constant under stricter or more lenient accuracy criteria?
5. Can an order of acquisition of different components be found for the individual words being tested?

I hypothesise that a systematic pattern of difficulty in the acquisition of word knowledge components can be expected, as it seems from general teaching experience that learners would, for example, acquire the form–meaning link of a word before they master all their derivative forms or meaning senses. This IS would allow us to check whether our intuitions of word knowledge order are statistically reliable.

Following the assumptions of implicational scaling, if it is found that the word knowledge components form an IS, it means that they belong to a one-dimensional continuum, can be ordered according to their difficulty and, thus, are expected to be acquired in that order. If it is indeed established that the different word knowledge aspects follow a pattern of acquisition, it would have strong implications for both vocabulary teaching and testing. It could inform teachers of which aspects of words to introduce to the students first and how to sequence the presentation of the remaining aspects. Regarding the testing field, if an order of acquisition is discovered, then we could make predictions about what types of word knowledge a student possessed of a word by demonstrating knowledge of another aspect. The next section describes the methodology implemented to answer the research questions.

4.3. Methodology

This study follows the research design described in Chapter 3. The battery of eight tests measuring recall and recognition knowledge of four vocabulary components were administered to the Spanish participants and the results analysed following the Guttman Scaling procedure.

4.3.1. Guttman Scaling procedure of goodness-of-fit

As stated above, the method of analysis used in this study is Implicational Scaling Analysis, aka Guttman Scaling. The first step in the application of this method is to create the implicational matrix, from which a pattern of difficulty of the linguistic features under examination is retrieved. However, implicational matrices are seldom perfect. In order to give weight to the evidence and establish whether this pattern of difficulty of items is valid we need to compute the statistical procedure developed by Guttman. This procedure checks the distribution of known/unknown features to

determine whether they can be ordered in terms of difficulty and to evaluate the goodness-of-fit of the scale.

The procedure involves computing different statistics: *minimum marginal reproducibility, coefficient of reproducibility,* and the *percent improvement,* with which we compute the *coefficient of scalability*. Following Hatch and Lazaraton (1991), I will proceed here to explain what these different statistics do:

1. *Minimum Marginal Reproducibility*

This figure informs us about how well we can predict the performance of a learner without taking into account the errors.

The formula for this is:

$$MMrep = \frac{\text{maximum marginals}}{(\text{number of learners}) * (\text{number of items})}$$

The maximum marginals refer to the number of correct/known responses to a particular item. In other words, how many participants demonstrated knowledge of a specific aspect. Ordering the linguistic items under study by their maximum marginals we can get an initial ideal about their difficulty. However, this coefficient does not take into account the number of errors (i.e., deviations from the predictive pattern), and thus the predictive power of this coefficient is less precise. The Coefficient of Reproducibility solves this problem.

2. *Coefficient of Reproducibility or Index of Reproducibility*

Errors occur even in valid ISs. These errors refer to the number of responses that deviate from the ideal pattern (i.e., if a difficult aspect is known, then all easier ones would have been mastered). Since the scale fit depends on the degree of errors, they need to be considered. In order to account for random errors and recover a scale from noisy data, the simplest method (called the Goodenough-Edwards method) is to order the rows and the columns according to their total sum of correctly answered questions, and then count the number of errors (Abdi, 2010). With this information, we compute the Coefficient of Reproducibility. This coefficient is an estimate of the implicational prediction in a scaling analysis (Guttman, 1944). That is, it informs us of how well we can predict a learner´s performance from their location in the scalogram. The ability to

predict a person's performance is more accurate when errors are taken into account, and thus, the Coefficient of Reproducibility value (C_{rep}) should be higher than the Minimum Marginal Reproducibility coefficient (MM_{rep}).

C_{rep} takes the total number of errors and divides this by the total number of possible errors (number of columns*number of rows). The formula for this coefficient is:

$$Crep = 1 - \frac{\text{number of errors}}{(\text{number of learners}) * (\text{number of items})}$$

A scale is deemed acceptable if it includes less than 10% of errors (Guttman, 1944; Van Schuur, 2011). That is, if the C_{rep} is above .90. Thus, by convention, the C_{rep} should be over .90 for a scale to be considered valid. Some researchers claim that we should establish .93 as the minimum C_{rep} figure, since Guttman (1944) stated that a C_{rep} value of .93 is approximately at the .05 significance level (Dunn-Rankin, 1983; Rickford, 2002). Following these values, a C_{rep} of .93 would mean that 93% of the time we could accurately predict which linguistic features a person answered correctly by their rank in the matrix. However, all previous evaluations of implicational data have adopted the minimum reproducibility value of .90 for a scale to be considered valid, and thus this will be the value followed in this study for comparability reasons.

3. *Percent Improvement in Reproducibility*

It indicates how much improvement there is between the C_{rep} and the MM_{rep}. This value is a stage for computing the final scalability coefficient. Thus, the formula is a simple subtraction:

$$\%\ improvement = Crep - MMrep$$

4. *Coefficient of Scalability*

It is the figure that shows whether a set of items are scalable, and thus, unidimensional. That is, whether the IS represents the data appropriately. It employs the following formula:

$$Cscal = \frac{\%\ \text{improvement}}{1 - \text{MMrep}}$$

For the data to be considered perfectly scalable, the Coefficient of Scalability (C_{scal}) should be 1. However, since a perfect scale is very unlikely, statisticians have determined that the C_{scal} must be above .60 to be able to claim scalability (Hatch & Lazaraton, 1991). The higher the value of the Coefficient of Scalability, the more "implicational" the scale is (Davidson, 1987). Therefore, even if the scalogram does not present perfectly scaled data, if a strong pattern of scalability is shown (C_{scal} >.60), we can claim that the data is scalable.

The two values that need to be reported in an implicational scaling study are the Coefficient of Reproducibility and the Coefficient of Scalability. The former shows how accurately the performance of a participant can be predicted given their position in the scalogram, and should be above .90 to consider a scale valid. The latter indicates whether the set of items tested are scalable and unidimensional, and ascertains the strength of those items, requiring the value to be above .60 for the data to be considered scalable.

4.3.2. Cut-off criteria for Implicational Scaling

In order to create an IS that establishes patterns of difficulty within linguistic items, such items need to be coded binary as known/mastered or unknown/not mastered. For that, decisions need to be made regarding the cut-off criterion used to consider an item known. Deciding on this cut-off point is an important stage in implicational scaling which requires careful consideration, as the placement of this threshold can affect the scaling results (Anderson, 1978). Although there is no specific guidance on the most appropriate threshold, previous research employing implicational scaling to study SLA has adopted the cut off of 80% accuracy for a particular linguistic item to be considered known (e.g., Andersen, 1978; Anderson, 1978; Ellis, 2008; Gitsaki, 1996; Newman, 2010). That is, if a person knows 80 to 100% of the items, it is considered that they have mastered that linguistic aspect, and thus given a 1. Accuracy of less than 80% is considered as not mastered and coded 0.

Because implicational scaling has not been extensively applied in vocabulary research, there is no clear answer as to whether this 80% criterion works for lexical items. However, Gitsaki (1996) applied the 80% accuracy criterion to the study of collocation acquisition with good results, which suggests that this criterion seems to be suitable to lexical items. Therefore, following previous linguistic research on

implicational scaling, I implemented the 80% accuracy criterion, which requires the participants to provide at least 80% correct answers for a particular word knowledge aspect to be considered mastered. However, although normally used in implicational scaling research, this level is arbitrary, and we cannot be completely confident about its suitability.

Since my study is exploratory, I did not consider it appropriate to apply only a single criterion. Thus, I tried two different, but reasonable, cut-off-point criteria to recognise an aspect as being known. Using a cut-off point higher than 80% to accept an aspect as known was considered an unreasonably difficult target. Similarly, 70% or less correct answers was not considered to be mastery of an aspect. Therefore, the cut-off criterion of 75% accuracy was also tested in this study. This is what I believe to be the minimum for a particular component to be considered acquired. Thus, the study checked scalability of the word knowledge components at the 80% and 75% accuracy levels. If a scale holds when employing different accuracy level criteria, this scale will gain statistical power.

4.3.3. Participants

Participants in this study were Spanish learners of English as a second language. The use of implicational scaling requires that the linguistic items tested have different degrees of difficulty, and that the subjects assessed have different proficiency levels. Therefore, I aimed for a sample of EFL learners with a range of proficiencies in English, from beginners to advanced.

Initially 148 Spanish learners of L2 English participated in this study. Some of them were still undergraduate or postgraduate students undertaking various degrees, while others were professionals of a range of disciplines. They were recruited as volunteers in Spain and the UK with the only requirements of being native speakers of Spanish and having some knowledge of English. Data from four participants was excluded from further analysis for two main reasons: 1) not having completed all the sections of the test, and 2) having been caught checking the answers on their phone during some of the sections. Thus, the final pool of participants consisted of 144 subjects (102 females (70.8%) and 42 males (29.2%)), who ranged in age 18-65 (M = 25.25, SD = 8.04).

The minimum amount of time the participants had studied English (in the form of formal instruction or self-instruction) was four years, and the maximum was 26 years. In fact, most of the participants (64.6%, 93 people) were studying English formally at the time of the study, either at university in order to get a certificate in English philology or education, or in academies or language schools with the purpose of obtaining a certificate of English proficiency (e.g., FCE or CAE). Almost a third of the participants (28%) reported having a degree in English (e.g., English teacher education, English Philology, Translation studies, etc.) at the time the study was conducted, and 43% of the remaining 103 participants had studied part of their degree in English as language of instruction or were working towards getting a degree in English. This suggests that a majority of the participants felt fairly confident about their English proficiency. In fact, almost half of the participants 49.3% (71) reported having an intermediate general proficiency level in English, and 38.5% (54) considered themselves advanced users of English, while only 13.2% (19 people) rated themselves as beginners in English. Yet, only 29.2% (42) of the participants had spent a minimum of three months in English speaking countries, with one participant working nine years in the UK, and 22 people spending one or more years doing a postgraduate degree in the UK. Finally, their general linguistic knowledge and learning skills seem substantial, as 60.6% of the participants reported speaking at least another second language apart from English.

Table 8 summarises the participants' overall exposure to English from different sources and modes. Listening to radio/music seems the most popular means of English input, followed by watching TV and reading.

Table 8 Spanish participants' exposure to English (per week).

	0 hours	**0-1 hours**	**1-2 hours**	**2+ hours**
Reading	10.4%	27.1%	25%	37.5%
Watching TV	12.5%	22.2%	24.3%	41.0%
Listening to Radio/Music	1.4%	13.2%	21.5%	63.9%
Using social networks to communicate	31.3%	31.3%	16.0%	21.5%
Playing computer/online games	65.6%	18.3%	8.4%	9%

In sum, this information suggests that, although participants had a range of proficiency levels, they had a fair amount of contact with English.

4.3.4. Test-battery administration

The data collection took place in Spain and the UK for 1.5 months. The test was administered in pen-and paper format, and participants handed in each individual test after finishing it and before starting the next one. The different tests were collected by the researcher one by one after completion by the participant. Each participant had an individual code that connected all their individual tests together.

After explaining the study orally and allowing time to individually read the information sheet, the participants signed the informed consent form (see Appendix 4 for the Participant Information sheet and a sample of signed consent forms[9], both documents provided in English or Spanish depending on the participants' individual preferences). They were allowed to ask questions at any time during the testing process. After completion of the whole test-battery, participants were asked to fill in the language background questionnaire. The whole test took about 2.5-3.5 hours to complete. All the participants were given £6 for their participation. Also, a detailed report of their results in each independent test was provided to those participants who requested it, in exchange for their help.

Although the participants were given the option to complete the test in two separate sessions to address their busy schedules (see Chapter 3, Section 3.3.3), only ten participants took this option, and the rest of them completed the test in one sitting. I believe the results are comparable since all participants had a long break in the same location in the test, and no significant difference was found in the results of the two groups.

Finally, the study employed a counter-balanced design. Participants were randomly assigned to one of the two versions of the test (the words in the first half of Version 1 were included in the second half in Version 2, and vice versa). Moreover, within those two versions participants were given the option to choose from the test with English or Spanish instructions and explanations, depending on their confidence with the language. 75 participants took Version 1 of the test (*N*=10 Spanish Version 1; *N*=65 English Version 1), while 69 completed Version 2 of the test (*N*=10 Spanish Version 2; *N*=59 English Version 2).

[9] The full set of consent forms by participants are available upon request.

4.3.5. Data analysis

The main method of analysis for this study is implicational scaling. Since this scaling technique was designed for categorical data, in order to create the implicational matrix I had to code each participant as having mastered/knowing (1), or not (0), the eight word knowledge aspects. To do this, I computed the total percentage of right answers across all the target words in each of the eight word knowledge aspects (i.e., how many words they knew in each aspects). Then, with this data, I applied the two different cut-off points to consider when a participant had mastered each aspect. For example, following the 80% accuracy criteria, if a participant had over 80% of correct answers in one aspect, I would consider the aspect known, and thus give them a 1. Aspects were arranged horizontally from left to right according to how well they were known (i.e., the number of participants that achieved the minimum score to consider mastery of that aspect), so that the best known aspect was on the left. Participants were vertically rank-ordered according to their general performance, (i.e., how many aspects they mastered), so that the participant with higher score overall was placed at the top.

The resulting data matrix was tested against the null hypothesis that claims no systematic order in the acquisition of word knowledge components in English by L2 learners of English. For each scale, the Guttman's coefficients of reproducibility (C_{rep}) and scalability (C_{scal}) were computed in order to test the validity of the scales and the scalability of the items in the scales. The aim is to explore whether the scale is highly reproducible (i.e. C_{rep} >.90) and highly scalable (i.e., C_{scal} >.60).

Guttman scaling is a deterministic model, and therefore, follows an unrealistic assumption that the data is error-free. Although the coefficient of scalability was designed to control for the possibility of errors, it still has limitations in this respect (Van Schuur, 2003). Therefore, I employed Mokken scaling (the probabilistic version of Guttman scales) to confirm and support the implicational scaling results. The Mokken technique is a non-parametric, confirmatory tool that uses statistics to describe unidimensional ordering. That is, whether a set of items are homogeneous, scalable and belong to the same underlying, latent construct. It extends the deterministic Guttman scaling model allowing the researcher to control for measurement error (Van Schuur, 2011). Mokken scaling provides not only a difficulty scale of linguistic features, but also measures its homogeneity (Scale *H*) and its reliability (*Rho*). Scale *H* is a value that

tells us whether a set of items form one dimension, and the *Rho* value is a score that measures the reliability of the scale. The *H* value should be above .30 for a scale to be considered valid. *H* coefficient between .30 and .40 is considered a weak scale, .40 to .50 is medium strength, and *H* >.50 indicates a strong scale (Sijtsma & Molenaar, 2002). As for the *Rho* value, it should normally be above .70 for a scale to be considered reliable (Sijtsma & Molenaar, 1987). If these coefficients are high, the interpretation is that the items can be hierarchically arranged in order of difficulty and form part of a unique construct. Therefore, it has been suggested that Mokken scaling can be applied as a secondary analysis to assist and check the performance of the Guttman scale (Stochl et al., 2012), as well as to complement the results by establishing a value for the dimensionality of a scale and for its reliability.

After collecting, preparing and entering the data following the scoring protocols described in Chapter 3 (Section 3.2), preliminary data screening and cleaning were carried out to check for errors and correct them, and the normality of the variables assessed to decide on the type of tests to use for the general analysis.

4.4. Results and interpretation

The presentation of results will be divided into general results and the implicational scaling results. The data for the present study was analysed using SPSS version 24 (IBM Corp., 2016) for general statistics and MSPWin5 (Molenaar & Sijtsma, 2000) for Mokken analysis.

4.4.1. General results

4.4.1.1. Reliability of the test battery

The reliability of each task in the test battery was examined by means of an estimate of internal consistency, measured by Cronbach's Alpha reliability value[10]

[10] Cronbach Alpha is the standard metric of reporting reliability in language measurement in general, and thus is the one I will report in this thesis. Yet, it may not be the best reliability measure for vocabulary studies, as each word is its own construct and knowing one does not imply knowing others. More sophisticated reliability measures, like the ones derived from Structural Equation Modelling analysis, would be better for vocabulary instruments. Hence, the final decision on the reliability of these tests will be based also on these other measures reported in Chapters 6 and 8.

(number of participants N = 144; number of items in each test k = 20). The reliability levels were good for all the tests (see Table 9).

Table 9 Internal consistency reliability for word knowledge aspects' tests (Spanish group).

Test	Cronbach Alpha (α)
Form–Meaning link Recall	α = .81
Form–Meaning link Recognition	α = .74
Derivative Recall	α = .95
Derivative Recognition	α = .94
Multiple-Meanings Recall	α = .91
Multiple-Meanings Recognition	α = .88
Collocation Recall	α = .80
Collocation Recognition	α = .80

Following the rule of thumb that the Cronbach's Alpha coefficient should be above .70 (DeVellis, 2012), we can consider the tests to be reliably assessing the aspects in question. Table 9 shows that the Derivative tasks are the most consistent tests, while the Form–Meaning Recognition test seems to be the least consistent one. This suggests that the latter test accounts for slightly more variability in the answers from participants than the rest of tests, and thus, exhibits less consistency in measuring the concept in question.

The Multiple-Meanings Recall test showed good consistency in measuring the specific aspect. Yet, this test requires open-ended answers by the participants, and thus it is liable to subjectivity in the scoring. To assess possible inconsistency due to scoring these open questions, two additional independent raters (apart from the researcher) proficient in both English and Spanish were trained and given detail scoring guides to score a random 20% sample (n= 30) of the total number of multiple-meanings recall tests analysed. Each test was assessed by each of the three raters.

When analysing inter-rater reliability, usually the Cohen's Kappa coefficient is reported. However, this analysis can only be done when the data is categorical. Since the total test score is a continuous variable, I computed the Intra-class correlation coefficient (ICC), a value used to check the reliability of inter-raters when the variable

is continuous. Because the reliability is calculated from a single measurement, not an average of the rater's measurements, I report the single measure coefficients.

A high degree of reliability was found between the three raters. The ICC across the three inter-raters was .92, with a 95% confidence interval from .86 to .96 ($F(29,58)=36.031$, $p<.001$). When the ICC reliability was tested between the researcher and each of the raters, the results remained strong. Researcher vs. Rater 1, the ICC coefficient was .95 with a 95% confidence interval from .90 to .98 ($F(29,29)= 38.343$, $p<.001$). The ICC coefficient between the main Researcher and Rater 2 was .91 with a 95% confidence interval from .83 to .96 ($F(29,29)=22.340$, $p<.001$). Finally, the ICC between Rater 1 and Rater 2 was .91 with a 95% confidence interval from .81 to .95 ($F(29,29)=19.772$, $p<.001$). In sum, the inter-rater ICC reliability for this test showed a general agreement of 92% between the three raters.

Overall, the tests showed to be reliable and demonstrated consistency in assessing the specific word knowledge aspects under consideration.

4.4.1.2. Normality check

Given the large sample in my study (>50), a Kolmogorov-Smirnov test with Lilliefors significance correction was used to check for normality of the distributions of the different tests' scores. Skewness and Kurtosis values, as well as histograms, were checked to further decide on the non-normality of the data. Only the Multiple-Meanings Recall test showed to be normally distributed (Multiple-Meanings Recall $D(144) = .053$, $p=.200$), with the rest of the tests' scores deviating from normality (VLT Total $D(144) = .096$, $p=.002$, Form–Meaning Recall $D(144) = .091$, $p=.006$, Form–Meaning Recognition $D(144) = .185$, $p=.000$, Derivative Recall $D(144) = .80$, $p=.026$, Derivative Recognition $D(144) = .105$, $p=.001$, Multiple-Meanings Recognition $D(144) = .080$, $p=.025$, Collocation Recall $D(144) = .090$, $p=.006$, Collocation Recognition $D(144) = .196$, $p=.000$). Therefore, non-parametric tests were employed.

4.4.1.3. Knowledge of the vocabulary components

To get a general estimation of knowledge of the different word knowledge components from the sample as a whole, descriptive statistics were computed. Table 10 shows the raw scores for each individual test.

Table 10 Descriptive statistics of the Spanish participants' performance (raw total scores).

Measure	**Min.**	**Max.**	**Max. Possible**	**Mean**	***SD***
VLT TOTAL	32	119	120	81.37	20.15
VLT 2K	8	30	30	24.86	5.48
VLT 3K	2	30	30	22.28	6.08
VLT 5K	8	30	30	21.38	5.38
VLT 10K	1	29	30	12.85	5.46
Form–meaning Recall	2	19	20	10.72	3.78
Form–meaning Recognition	8	20	20	16.56	2.76
Derivative Recall	4	73	80	41.31	16.69
Derivative Recognition	15	78	80	48.86	15.92
Multiple-Meanings Recall	4	54	60	29.81	9.63
Multiple-Meanings Recognition	19	58	60	42.46	9.11
Collocation Recall	3	19	20	12.13	3.81
Collocation Recognition	5	20	20	15.94	3.48

Note: $N = 144$

Since each word knowledge component had a different scoring system and maximum score, to make the information more interpretable, the descriptive statistics are presented in percentages in Table 11.

Table 11 Descriptive statistics of the Spanish participants' performance (percentage total scores).

Measures	Min. %	Max. %	Mean %	*SD*
VLT TOTAL	26.67	99.17	67.81	16.79
VLT 2K	26.67	100	82.87	18.28
VLT 3K	6.67	100	74.26	20.26
VLT 5K	26.67	100	71.25	17.93
VLT 10K	3.33	96.67	42.85	18.21
Form–meaning Recall	10	95	53.61	18.91
Form–meaning Recognition	40	100	82.81	13.79
Derivative Recall	5	91.25	51.64	20.86
Derivative Recognition	18.75	97.50	61.08	19.90
Multiple-Meanings Recall	6.67	90	49.69	16.12
Multiple-Meanings Recognition	15	95	60.66	19.05
Collocation Recall	31.67	96.67	70.76	15.18
Collocation Recognition	25	100	79.69	17.39

Note: $N = 144$

The descriptive statistics of the word knowledge tests show that learners knew, on average, at least half of the items in each test, with the recognition mastery of each aspect showing a higher mean score than its recall mastery. Moreover, the best-known aspect on average is, unsurprisingly, form–meaning recognition, while knowledge of multiple meanings and derivatives overall seem to be particularly complex for learners. The results also show that the participant population was appropriately diverse, producing a range of results (i.e. a range of weaker, mid, and stronger scores), which was necessary for the implicational scaling procedure.

In the case of derivatives, learners appeared to be confused regarding the suffixes needed to create different word classes. Some instances were found where learners interpreted a word with the suffix *–ness* as an adjective, or *–ly* as a noun or adjective, and *-ing* as adverb (e.g., *brightness* as adjective, *bright* as noun and verb, and *brighten* as adjective; *freshly* as adjective, or *challenging* as adverb). This suggests that learners do not seem to have complete distributional knowledge of derivatives (i.e., the

form-function knowledge), which would explain why they know some suffixes, but not how to use them appropriately, thus applying them to the wrong word class. Moreover, on some occasions a learner answered a word correctly in the derivatives recall section and incorrectly in the recognition section. This may be explained by the fact that even if they can produce the derivatives of words relatively well, when encountering multiple form options that follow the morphological rules they appear to get confused. An example of this is producing the word *hard* as an adjective in the recall derivative section and then choosing the distractor *hardful* for the adjective in the recognition derivative test. These situations indicate that participants generally know the suffixes and morphological rules used to create different word forms (i.e., relational knowledge), but may not know which suffix is attached to each form or if it is the best or only option (i.e., distributional knowledge).

Multiple meanings seems to be another aspect that caused problems for participants. It was found that learners seem to have some partial knowledge of the multiple-meanings recognition aspect, even if they did not know the exact meaning or equivalent. For example, they could know the meaning of 'cooking style' for the target word *grate* when it was used in a sentence, but when asked to recall that meaning in the multiple-meanings recall test it was defined as 'to grill'. This suggests that they had some general knowledge of the meaning of this word being related to cooking or food, but they did not know the exact meaning of the word, which did not generally prevent them from understanding it. Also, as expected, multiple-meaning words in which the different meaning senses are completely different from one another (i.e., homonyms, such as *bank* or *season*) are worse known than those words whose multiple meanings are very similar among them (i.e., polysemes, such as *hard*) or very similar to the Spanish ones (e.g., *terminal*).

However, there are also some unexpected results. In particular, the mean score for the Collocation Recall test proved to be higher than that of the Form–Meaning Recall test. I hypothesise that the reason is the similarity in test formats between both tests, combined with the frequency of the target words required to be recalled. In the Form–Meaning Recall test the target words tested came from a range of frequency levels, some as low as 9K. The Collocation Recall test required the students to recall only one of the target word's collocates, which belonged to the first 3K frequency bands and for which the first letter was provided. This format of testing recall knowledge of

collocations has been shown to work for collocations of various frequencies and when both constituents of the collocation are required to be produced (see González Fernández & Schmitt, 2015). Yet, it may be that measuring very frequent collocates when presenting the target word results in a slightly easier task than requiring demonstrations of Form–meaning Recall.

Regarding the Vocabulary Levels Test (VLT), I scored the four sections assessed (2K, 3K, 5K and 10K) counting the number of correct answers out of the possible 30. The general, VLT Total score was computed adding up the scores from the four individual sections, and thus it is scored out of 120 (30 words x 4 tests). Although this test was not designed to sum up all the scores in the different sections, I decided to take this approach in order to obtain an estimate of the participant's general vocabulary proficiency, yet understanding that this score does not represent the whole proficiency level of participants.

From the VLT scores in Table 11 we can see that participants knew on average 68% of the words tested in the VLT (81.4/120 right answers), with 71% right answers in 5K level and 43% in 10K. This suggests that their overall vocabulary mastery is good. However, participants do not seem to have mastered any level (one suggested mastery score is answering correctly 28/30 (93.3%) words). They all demonstrate knowledge of some of the 10K level words, even if they have not mastered the previous levels. This is likely to reflect the cognate effect. This role of cognateness status of words will be explored and interpreted in light of the results of Study 2 (Chapter 5).

Although it seems possible that the mean scores in each test can outline a tentative order of acquisition of vocabulary components, mean scores can be misleading. Individual learners do not necessarily have to experience the same order of difficulty as the sample as a whole. For this reason, implicational scaling analysis is required to claim an order of acquisition, since it examines the patterns for each individual participant.

4.4.1.4. Correlations between word knowledge components and the VLT

To answer the research question on the extent to which the different word knowledge components are interrelated, correlational analyses were computed. Because

the main variables in this study were non-normally distributed, Spearman's correlations were calculated for the mean scores between the word knowledge component aspects.

Table 12 Correlations between word knowledge components (Spanish participants).

	VLT Total	FM Recall	FM Recog	Deriv Recall	Deriv Recog	MM Recall	Collo Recall	MM Recog
VLT Total								
FM Recall	.786**							
FM Recog	.760**	.773**						
Deriv Recall	.895**	.811**	.803**					
Deriv Recog	.865**	.786**	.828**	.945**				
MM Recall	.872**	.747**	.774**	.845**	.818**			
Collo Recall	.839**	.772**	.749**	.827**	.812**	.817**		
MM Recog	.826**	.720**	.766**	.835**	.831**	.848**	.771**	
Collo Recog	.779**	.678**	.700**	.763**	.768**	.752**	.806**	.737**

Spearman: **$p < .01$

VLT = Vocabulary Levels Test; FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.

Table 12 shows that the correlations between word knowledge components lie between .68 and .95, showing a high degree of interrelationship. These intercorrelations are all very large, according to Cohen (1988)[11].The Derivative Recall aspect seems to be the one that correlates more strongly with the rest of the aspects.

It is also interesting that knowledge of the various word knowledge components correlate very highly with their vocabulary size, as indicated by the VLT Total score. The highest correlation with the VLT is not the Form–Meaning Recognition (r =.76, p < .01), but Derivative Recall (r =.90, p < .01), followed by Multiple-Meanings Recall (r =.87, p < .01). This is not surprising because the Derivative Recall and Multiple-Meanings Recall tests were the most difficult aspects, and so only more advanced learners with larger vocabulary sizes were able to master these word knowledge components. Conversely, most learners (even ones with lower vocabulary sizes) were able to master the Form–Meaning Recognition aspect, so depressing the correlations.

[11] Cohen (1988, p.79-81) suggests the following guidelines to evaluate the strength of relationships: r =.10 to .29 represents a small interrelation, r =.30 to .49 a medium one, and r =.50 to 1 a large correlation.

The results support Schmitt's (2014) conclusion that vocabulary size and depth are inextricably related to one another, and that word knowledge components seem to be highly interconnected. It can also be seen that the correlations between VLT and the recall tests are slightly higher than for recognition tests, suggesting that as the general vocabulary level of the participants improves, so does the recognition and recall mastery of all aspects, although with a minimal advantage of recall mastery.

For each of the four general word knowledge components, the correlations between its recall and recognition masteries show the same general pattern of high relatedness. Form–Meaning Recognition and Recall correlated at .77, $p < .01$, Derivative Recall and Recognition at .95, $p < .01$, Multiple-Meanings Recognition and Recall at .85, $p < .01$, and Collocation Recall and Recognition at .81, $p < .01$. Finally, the recall aspects correlated between .75 and .85, ($p < .01$), and the recognition aspects between .70 and .83 ($p < .01$), suggesting they belong to the recall and recognition vocabulary knowledge concepts respectively.

4.4.1.5. Summary of general results

To sum up, the general results suggest that the word knowledge component learners know better is the form–meaning link overall, and that the components they had less knowledge of were knowledge of derivatives and multiple meanings. Recognition knowledge is also known better than recall knowledge. Despite these differences in difficulty, the word knowledge components seem to be very highly interrelated with each other, which suggest that they are all part of a higher concept.

4.4.2. Implicational scaling results

Having presented above the level of knowledge of the word knowledge aspects by the sample, in this section I will present the implicational scaling results, showing the order of acquisition of the different components as a whole and per word. The IS of the different vocabulary knowledge aspects was created with the data from the 144 participants on the knowledge of 20 words in the different aspects.

4.4.2.1. 80% accuracy criterion

To create this implicational matrix, participants were scaled on the basis that they supplied at least 80% correct answers per aspect. The two axes of the matrix consisted of the eight word knowledge items horizontally, ranked from better known to worse known, and the 144 participants vertically ranked in order of their general performance, from participants knowing the most word knowledge aspects to the least.

Following Rickford (2002), I first present the matrix with percentages of correct answers, before converting them into 1s and 0s. This percentages matrix provides an accurate description of the knowledge of the different aspects per person (see Appendix 5). After doing the implicational scaling calculation with this percentages matrix, and following the most traditional scaling conventions, I converted the values into 1s and 0s following the data analysis procedures to produce the final implicational matrix[12] (see Appendix 6). Both scalograms retrieved the same scale (see Figure 5).

Figure 5 Order retrieved in the implicational scale for the Spanish group with the 80% accuracy level (from easier to more difficult).

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

In order to find out whether the scale is valid, and if so, how good it is, I computed the Guttman statistics. The coefficients are presented in Table 13.

[12] In order to decide where to draw the line that divides mastery and lack of mastery in each aspect per participant, I took into consideration the previous and next aspect of knowledge, so that the previous level always included at least the same number of participants answering correctly as the next level.

Table 13 Results of Guttman Procedure with 80% accuracy criterion (Spanish group).

	C_{rep}	MM_{rep}	% improvement	C_{scal}
Whole matrix	**.94**	.78	.16	**.73**
Recognition part of matrix	**.93**	.68	.25	**.78**
Recall part of matrix	**.95**	.88	.07	**.60**

C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % improvement = Percentage of improvement in reproducibility; C_{scal} = Coefficient of scalability.

Following the conventions for the values of these statistics (see Section 4.3.1), C_{rep} must be over .90 before a scale can be considered valid, and C_{scal} above .60 before scalability can be claimed. Table 13 shows that the IS of word knowledge components based on the 80% accuracy criterion reaches the statistical coefficients to be considered valid, both as a whole and dividing it into a recognition and recall scale.

Given that the C_{rep} of the whole matrix satisfies the minimum reproducibility value of .90 (i.e., C_{rep}=.94), and is above the .93 value considered to be an approximation of a .05 significance level (Guttman, 1944), we can say that these word knowledge components form a valid IS. This means that knowledge of a higher aspect in the scale implies knowledge of all lower aspects. Thus, according to this C_{rep} if a participant knows one aspect in recall it would mean that around 94% of the time that person would know all the aspects in recognition receptively. The high C_{scal} (.73) indicates that we can be confident that an IS exists, and that the pattern of scalability is very robust. Thus, we can conclude that the word knowledge components are scalable and the implicational relationship between them is significant and strong (see Andersen, 1978; Hatch & Lazaraton, 1991).

In order to confirm the results from the IS, I ran a Mokken scaling analysis employing the same 1-0 matrix for the 80% accuracy criterion. Table 14 shows the results.

Table 14 Mokken scale for 80% accuracy criterion, reliability and dimensionality assessment (Spanish group).

Order	Aspect	Mean	Item H
1.	Form–meaning Recognition	.67	.87
2.	Collocation Recognition	.63	.90
3.	Multiple-Meanings Recognition	.35	.78
4.	Derivative Recognition	.24	.78
5.	Collocation Recall	.22	.68
6.	Form–meaning Recall	.11	.70
7.	Derivative Recall	.10	.81
8.	Multiple-Meanings recall	.05	.85

Scale *H* = .79

Rho = .89

The results from the Mokken analysis show that the order of difficulty of the tests is the same as suggested in the IS presented in Figure 5, and thus, supports the previous scale. The homogeneity of the scale is very strong (Scale *H* = .79), indicating that the aspects are highly intercorrelated and form part of an underlying construct (i.e., vocabulary knowledge). The reliability of this scale is also very high (*rho* = .89). Besides, none of the individual aspects fall below the recommended value of *H* >.3 to be considered unidimensional. Indeed, considering their *H* values, the aspects could all be considered strongly unidimensional, suggesting that they all are essential items of the same scale.

Given the evidence that the scale is valid and reliable, we can start analysing the difficulty pattern. Regarding the order of the aspects in the scale, we can see a very neat division between recognition and recall knowledge. This means that recognition mastery seems to be easier than the recall mastery in all aspects.

The difference in knowledge between recognition and recall mastery overall is logical, and represents the generally agreed notion that recognition knowledge is acquired before recall knowledge. However, the fact that recognition knowledge of *all* aspects seems to be mastered before the recall knowledge is somewhat surprising. It could have been expected that in easier aspects, such as the form–meaning link, the recognition mastery would be followed by the acquisition of its recall mastery, before recognition knowledge of any other component is achieved. However, the implicational

scaling and Mokken results suggest that overall recognition knowledge seems to be easier than any recall knowledge.

Inspection of the recall section reveals the same surprising finding suggested by the mean scores in Table 11. According to the scale, knowledge of recall collocations seems to be easier than recall knowledge of the form–meaning link. As described in Section 4.4.1.3, this unexpected result is partially explained by the similarity between both test formats and the difference in the frequency of the words needed to be recalled. In the form–meaning link recall test, learners needed to spell the target word (from 1-9K frequency bands), while in the collocation recall test they needed to provide the collocate, all of which came from within the first 3K frequency bands. It appears that, overall, learners know the very frequent collocates to a form-recall mastery marginally better than they know the spelling of all the lower-frequency target words. Combing the target words of various frequencies, the general acquisition order tipped in favour of recall collocation being somewhat easier than recall form–meaning overall. Thus, this seems to reflect that the partial recall collocational knowledge represented by the test is easier to master than the full recall knowledge of form given the meaning, aspect that has been previously found to be particularly difficult (Laufer & Goldstein, 2004).

Overall, the application of the 80% accuracy criterion produces an IS of the various word knowledge components that is statistically and theoretically valid. The order of difficulty of components in the scale shows that recognition knowledge is easier to learn than any recall knowledge, that the form–meaning link seems to be acquired earlier, and that knowledge of collocations is less difficult for students than knowledge of multiple meanings or derivatives.

4.4.2.2. 75% accuracy criterion

To check that the results are not just an artefact of the 80% accuracy criterion, I applied the 75% accuracy criterion to code the components as known. Appendices 7 and 8 show the scalogram for the word knowledge aspects following the 75% cut-off level, in percentages and binary respectively.

As it can be seen in Figure 6, the IS suggested above holds regardless of the cut-off criteria used.

Figure 6 Order retrieved in the implicational scale for the Spanish group with the 75% accuracy level (from easier to more difficult).

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

In order to find out whether the scale continues to be valid with this cut-off point, and if so, how good it is, the Guttman statistics were performed. The coefficients are presented in Table 15.

Table 15 Results of Guttman Procedure with 75% accuracy criterion (Spanish group).

	C_{rep}	MM_{rep}	% improvement	C_{scal}
Whole matrix	**.95**	.74	.21	**.81**
Recognition part of matrix	**.95**	.66	.29	**.85**
Recall part of matrix	**.94**	.83	.11	**.65**

C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % improvement = Percentage of improvement in reproducibility; C_{scal} = Coefficient of scalability.

Again, the scale satisfies the minimum C_{rep} and C_{scal} values for it to be considered valid and robust. In fact, the scalability of the resulting data with the 75% accuracy criterion yielded slightly better coefficients than the 80% cut-off level (C_{scal} .81 vs. C_{scal} .73). However, the difference is minimal, and since both thresholds retrieve the same scale, I would deem the 80% accuracy criterion to be the most appropriate to consider mastery of a word knowledge aspect. Also, as in the previous analysis, the recognition and the recall parts of the scale independently also conform to an IS.

The Mokken analysis for the data with the 75% accuracy criterion adopted produced also very good results in terms of homogeneity and reliability of the scale (see Table 16).

Table 16 Mokken scale for 75% accuracy criterion, reliability and dimensionality assessment (Spanish group).

Order	Aspect	Mean	Item H
1.	Form–meaning Recognition	.74	.93
2.	Collocation Recognition	.69	.92
3.	Multiple-Meanings Recognition	.44	.88
4.	Derivative Recognition	.35	.88
5.	Collocation Recall	.26	.75
6.	Form–meaning Recall	.19	.77
7.	Derivative Recall	.17	.84
8.	Multiple-Meanings Recall	.08	.91

Scale *H* = .85

Rho = .92

Overall, these results seem to suggest that the scale of difficulty of word knowledge components found is very robust, as it proves to be consistent and statistically reliable regardless of which of these two cut-off levels is used.

Finally, it should be made clear that although this IS indicates that the aspects are hierarchically related, showing that some aspects are more difficult than others, it does not provide information about the extent of each aspect´s difficulty. In fact, the layout of the implicational matrix may give the wrong impression that the differences in difficulty between the aspects are the same. However, it is likely that the difference in difficulty vary between different pairs of items. It has been suggested that one way to know the size of difference between two contiguous aspects is by considering that the difference is inversely proportional to the number of participants that mastered each aspect (i.e., an easy aspect will be mastered by a larger number of participants than a difficult one) (Abdi, 2010). Thus, looking at the total number of learners that mastered each aspects (i.e., total number of 1s in each aspect) we can estimate the size of the difference in difficulty between aspects. Table 17 presents the total number of participants who mastered each aspect applying both accuracy criteria.

Table 17 Number of participants mastering each aspect (Spanish group).

	FM Recog	**Collo Recog**	**MM Recog**	**Deriv Recog**	**Collo Recall**	**FM Recall**	**Deriv Recall**	**MM Recall**
N of participants mastering 80%	97	91	50	34	32	16	15	7
N of participants mastering 75%	107	100	63	50	38	27	25	11

Table 17 shows that, although Form–Meaning Recognition was easier for participants than the Collocation Recognition test, they did not differ much in their difficulty. Conversely, Collocation Recognition seems to be clearly easier than Multiple-Meanings Recognition, and the latter in turn considerably easier than the next one, Derivative Recognition. The most difficult recognition aspect, Derivative Recognition, does not seem more different in difficulty than the easiest recall aspect, Collocation Recall considering 80% accuracy, although it seems different when the 75% cut-off is applied. Surprisingly, Form–meaning Recall seems considerably more difficult than Collocation Recall, but similar in difficulty to Derivative Recall. Finally, there seems to be an important difference in difficulty between the two most difficult aspects, Derivative Recall and Multiple-Meanings Recall.

Overall, we can conclude that there are different degrees of difficulty in mastering the different aspects along the scale. The order of difficulty and acquisition of the word knowledge components is clear and robust, but the difference in difficulty between them varies.

4.4.2.3. Implicational scaling per word

The implicational scaling analysis of word knowledge aspects provides information about how each aspect is likely to be acquired. However, it does not reveal information about the individual words and potential similarities/differences in their acquisition.

In the design of the study, I decided to test the same target words across all aspects, which allows me to explore how each of the 20 words is learnt across all aspects. Consequently, I checked whether the pattern of difficulty found in the general

IS holds for each individual word. My hypothesis is that each word would follow a different order of difficulty in the aspects, since some words have easier collocations than multiple meanings, or easier derivatives, or forms than others. In fact, in the selection of stimuli these differences were already found. However, it could also be that a majority of words follow the same pattern found above, providing stronger support for the general scale.

To investigate this, I computed the scalogram for each of the 20 target words in all aspects, arranging the order of aspects left to right from easier to more difficult (i.e., the aspects with more correct answers to the ones with less correct answers), and later arranging the participants top to bottom from more to less correct answers.

Table 18 shows the order of aspects' difficulty per word and the information about the Guttman's calculations.

Table 18 Results of Guttman Procedure for individual words (80% accuracy) (Spanish group).

Target word	Order of Aspects	C_{rep}	MM_{rep}	% $_{improve}$	C_{scal}
Close	MMRg>FMRc>CRc>CRg>FMRg>DRg>MMRc>DRc	**.91**	.83	.08	**.47**
Season	FMRg>FMRc>MMRg>CRg>DRg>DRc>MMRc>CRc	**.87**	.75	.12	**.48**
Bank	FMRc>FMRg>MMRg>CRg>CRc>DRg>DRc>MMRc	**.87**	.80	.07	**.35**
Terminal	FMRg>MMRg>MMRc>CRg>DRg>CRc>FMRc>DRc	**.87**	.77	.1	**.43**
Bright	FMRg>CRc>CRg>FMRc>MMRg>MMRc>DRg>DRc	**.91**	.80	.11	**.55**
Distinction	FMRg>CRg>MMRg>CRc>MMRc>DRc>DRg>FMRc	**.86**	.79	.07	**.33**
Draught	CRg>CRc>FMRg>MMRg>DRg>DRc>MMRc>FMRc	**.88**	.78	.1	**.45**
Challenge	FMRg>MMRg>CRg>FMRc>DRg>DRc>CRc>MMRc	**.88**	.70	.18	**.60**
Broad	CRg>FMRg>MMRg>CRc>MMRc>DRg>FMRc>DRc	**.87**	.65	.22	**.63**
Grate	CRg>CRc>MMRg>FMRg>DRg>DRc>MMRc>FMRc	**.90**	.78	.12	**.55**
Mean	FMRg>MMRg>FMRc>DRg>CRg>MMRc>DRc>CRc	**.86**	.71	.15	**.52**
Hard	FMRg>MMRg>MMRc>CRc>CRg>FMRc>DRg>DRc	**.92**	.81	.11	**.58**
Employ	FMRg>MMRg>CRg>CRc>DRg>MMRc>FMRc>DRc	**.85**	.74	.11	**.42**
Fresh	CRg>MMRg>MMRc>CRc>FMRg>DRg>FMRc>DRc	**.89**	.79	.1	**.48**
Charm	FMRg>MMRg>FMRc>CRg>DRc>DRg>MMRc>CRc	**.85**	.60	.25	**.63**
Indent	CRg>CRc>FMRg>DRc>DRg>MMRg>FMRc>MMRc	**.90**	.78	.12	**.55**
Fulfill	CRg>CRc>FMRg>MMRg>DRg>FMRc>DRc>MMRc	**.85**	.70	.15	**.50**
Character	FMRg>MMRg>MMRc>CRg>CRc>FMRc>DRg>DRc	**.89**	.83	.06	**.35**
Redeem	CRg>MMRg>DRg>FMRg>MMRc>CRc>DRc>FMRc	**.84**	.70	.14	**.47**
Development	FMRg>MMRg>FMRc>MMRc>CRg>DRg>DRc>CRc	**.88**	.74	.14	**.54**

For each aspect: *Rg = Recognition; Rc = Recall; C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % $_{improve}$ = Percentage of improvement in reproducibility ; C_{scal} = Coefficient of scalability*
Bold = C_{rep} and C_{scal}; Blue = approaching minimum value; Red = not reaching minimum value.

As seen in Table 18, none of the 20 words reach the minimum goodness-of-fit coefficients to consider the existence of an IS (C_{rep} of .90 and C_{scal} of .60). Most words do not achieve the minimum level in either of the two coefficients. There are a few words in which the coefficient of reproducibility is at the level necessary for the IS to be considered valid, but the coefficient of scalability is below the recommended level of .6 (e.g. *close*, *bright* or *hard*). This suggests that whilst the IS for the data is valid, the variance in terms of numbers of errors resulted in the low coefficient of scalability, and therefore we cannot claim an implicational relationship. Although three words (i.e., *challenge*, *broad*, and *charm*) seem to have achieved a valid C_{scal}, they did not reach a minimum C_{rep}, which is the first requirement for the existence of ISs. According to Guttman, less than 90% (.90) C_{rep} cannot be considered an adequate approximation to a perfect scale. Therefore, these words do not reach the minimum coefficient to suggest the possibility of an IS.

These results show that there does not seem to be a unique scale for individual words, since each of them follow a different pattern. Yet, those patterns do not reach a statistically reliable order, and therefore we cannot claim that there exists an IS for each individual word. The reason for these various results and lack of systematicity at the individual word level seems to be clear: a combination of variability within the words (i.e., each word has its own characteristics, with some being more difficult for certain aspects than others) and individual differences within the learners (i.e., some people may know some aspects of a word better than others for reasons not necessarily related to the intrinsic difficulty of those aspects, but the way it was taught, the exposure to that word in that aspect, etc.).

Finally, the lack of a valid scale for every target word suggest that the frequency of words or its cognateness status do not seem to be a factor in the likelihood of a word following an IS in the difficulty of its aspects. It may have been thought that high-frequency words may have followed a more varied acquisition pattern, as students would have more chances to encounter them in multiple occasions and varied contexts, whereas low-frequency words could have presented a more systematic pattern. Similarly, it may have been hypothesised that cognate words could follow a clearer pattern of difficulty in the acquisition of the different aspects, particularly having a sample from the same L1. In that sense, students may have been expected to know in a similar manner even the difficult aspects of cognate words that were analogous in their

L1. For example, the target word *fresh*, whose form is a cognate in Spanish, has multiple meanings which are very similar in Spanish, and thus even low-proficiency participants may have known that aspect, leading to a particular order of difficulty for that word. However, it may also be that those participants who apply the cognate rule to all aspects may be mistaken in the aspects that do not work similarly in both languages, and thus the variability of responses of participants would be greater, resulting in a lack of IS. Hence, the learners' individual differences seem to be a stronger factor influencing the order of acquisition of word knowledge aspects than frequency and cognateness.

Overall, we must conclude that, while an IS was found for the word knowledge aspects, there does not seem to be an IS at the individual word level. This indicates that the order of difficulty of word knowledge components is to be expected when we are measuring general knowledge of an aspect, but this order does not apply to each word individually. There are other features that make learning a word difficult, and thus affect their pattern of acquisition, such as the L1, the involvement with the L2, how the word has been learnt or taught and the specific internal complications of a particular word (Webb & Nation, 2017). Thus, whilst the sequence of acquisition of vocabulary components is robust, there still is a considerable amount of individual learner and item variation within each stage of vocabulary acquisition (Ellis, 1989).

4.5. Discussion

This study explored the order of acquisition of vocabulary knowledge components, using the most extensive multicomponent test battery to date (≈2.5-3.5 hours), with a large number of participants (N=144). This large amount of data confirmed, in some cases, common assumptions about vocabulary which have been made until now with little or no empirical evidence. In other cases, it led to surprising results, which challenge typical views of lexical knowledge.

In this section, the main results of the study will be presented separately, summarised, and interpreted in light of previous research. I will start by discussing the knowledge of the different vocabulary components shown by the participants, followed by the interrelationships between the components. Finally, I will consider the order of acquisition of such components, the influence of the two different cut-off criteria and the difference in difficulty between the contiguous aspects.

4.5.1. Knowledge of word knowledge components

Vocabulary knowledge is generally recognised as a complex notion that involves learning multiple aspects of a word (Nation, 2013). There is a general lack of understanding, however, about how well these aspects are known by L2 learners in comparison with each other.

The results from this study suggest that the participants possessed, as a whole, a generally good mastery of the different vocabulary knowledge components. They knew on average about 50% of the words in the recall aspects, and between 60% and 80% of the items in the recognition components. This finding confirms previous research demonstrating that the vocabulary recognition knowledge of L2 learners generally surpasses their recall knowledge (e.g., Webb, 2008).

It seems logical to expect that some of the word knowledge components are generally better known than others. My data confirms and provides evidence for this intuition. Recognition knowledge of the form–meaning link was, unsurprisingly, found to be the best known aspect, which is in line with previous research (e.g., Laufer & Goldstein, 2004). This link is generally the first contact a learner has with a word in the L2. Recognition knowledge of collocations was found to be known better than recognition of derivatives and multiple meanings, confirming the hypothesis that collocational recognition knowledge seldom causes as big of a problem for L2 learners as its recall knowledge (Henriksen, 2013; Laufer & Waldman, 2011; Nesselhauf, 2005). It has been suggested that the reason for this situation is that when learners know the component words, they can generally deduce the meaning of collocations even without recognising the collocation as a whole unit (Wolter & Yamashita, 2017). This finding contradicts the general notion that knowledge of collocations lags behind general language and vocabulary knowledge (Bahns & Eldaw, 1993; Irujo, 1993). If vocabulary knowledge is understood as the form–meaning link, collocational knowledge is poorer; when vocabulary knowledge is understood as mastery of other aspects, in this case derivatives and multiple meanings, then collocational knowledge seems to be better known. The results also provide further evidence for the difficulty of derivatives and multiple-meanings knowledge for L2 learners, which seem to pose the most difficulty for students both in the recognition and recall masteries.

The surprising finding that recall knowledge of collocations preceded recall knowledge of the form–meaning link highlights the difference that the testing format

makes in the evaluation of knowledge, as Kremmel and Schmitt (2016) concluded. When the node of the collocation (in this case the target word) is given and L2 learners are only required to provide a very frequent collocate of that node, they are able to make the necessary connections in the mental lexicon to recall the collocate. Nevertheless, when more complex collocations are assessed, and the learners are required to produce both terms of the collocation with limited context, they show poor collocational knowledge (Boers, Demecheleer, Coxhead, & Webb, 2014; Mizrahi, 2016; Qian, 2002; Webb, 2008). This finding may indicate that the latter collocational knowledge test format, although potentially a more accurate representation of free writing ability, is not sensitive enough to reflect partial recall knowledge of collocations. Conversely, the format employed in this study to assess collocational recall knowledge seems able to mirror initial recall knowledge of frequent collocations. At the same time, however, requiring only the retrieval of the high-frequency collocate in the collocation recall test made this test easier than the form–meaning recall test which required recalling lower-frequency words.

Overall, the findings regarding knowledge of components confirm previous studies and provide an insight into how the knowledge of some components relate to the knowledge of others. First, the form–meaning link was found to be the best known by L2 learners, particularly in its recognition form. Second, knowledge of derivatives and multiple meanings proved to be the most challenging for students. Finally, recognition mastery was found to be easier than recall mastery across all components, suggesting that the recognition/recall knowledge may be better seen as a continuum in which recognition mastery of all aspects is better than any recall mastery, even the form–meaning recall link.

4.5.2. The interrelatedness of word knowledge components

Ever since Nation first presented his initial listing of eight word knowledge components in his 1990 book, people have thought about the components as separate entities. Nevertheless, there have been claims that the components are interrelated in some way (e.g., Qian, 2002; Webb, 2008). It is also probable that many/most researchers and practitioners would assume some sort of connection between the components. My data demonstrates that these claims and intuitions were correct. I found

significant correlations between the various aspects ranging from .700 to .945, indicating that there are clearly strong and comprehensive connections between all of the word knowledge aspects I measured. This is interesting because it seems that no aspect is learnt in a way that is detached from the other aspects. Rather, the various components seem to be learned in a relatively parallel manner, as Webb (2007a) concluded. This suggests to me that knowing one aspect facilitates the learning of others.

There has also been a great deal of debate concerning the relationship between breadth (size) and depth (quality) of vocabulary knowledge. Depth has been conceptualised in many ways (Schmitt, 2014), and my study adds to this discussion by demonstrating the close relationship between vocabulary size (as measured by the VLT) and the various word knowledge components. The correlations are closely clustered between .760 and .895. This means that the shared variance between vocabulary size and word knowledge 'depth' ranges between 58% and 80%. Employing a form-recall size measure, Chui (2006) also found strong correlations (ranging between .53 and .78) between size and five word knowledge aspects, including derivatives, collocations, and word class. The VLT measures knowledge of the form–meaning link at the 'form recognition' level (Schmitt, 2010), so one might assume that the VLT scores would correlate more closely with the recognition and recall form–meaning tests. I found that the VLT actually correlates more strongly with the derivatives, multiple-meanings, and collocation measures. This suggests that vocabulary size is strongly related to knowledge of the various word knowledge components in general, and not just the ones which have a similar testing format (i.e., focusing on the form–meaning link). There is also a tendency for the size-aspect correlations to be somewhat higher for the recall measures than the recognition measures. This supports Chui's (2006) findings that size correlates more strongly with more difficult contextual aspects (i.e., derivatives or collocations) and the productive use of a word (i.e., sentence production). Overall, it appears that the greater your vocabulary size, the better you will know the words across all aspects.

4.5.3. Order of acquisition of word knowledge components

Although the various word knowledge aspects are closely related and somewhat acquired in a relatively parallel manner, they are not mastered completely simultaneously. Intuitively, it is reasonable to believe that some components would clearly seem easier (e.g. form–meaning link) or more difficult (e.g. multiple meanings) than others, and thus are expected to be acquired earlier. But until now, there has been little hard evidence indicating the order of acquisition of the various components. My study is the first to my knowledge to demonstrate a consistent and statistically-reliable IS of vocabulary knowledge components (see below).

Figure 7 Implicational scale of word knowledge components.

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

The findings from the IS can be divided into two main points of discussion: recognition vs. recall knowledge, and order of individual components.

4.5.3.1. Recognition vs. Recall

Previous studies focusing only on single word knowledge components have shown that recall knowledge seems to be generally more difficult than recognition for an individual aspect (e.g., the form–meaning link (Laufer & Goldstein, 2004), collocations (Peters, 2016)). But previous research was inconclusive as to whether all recall aspects are more difficult than all recognition aspects, or whether some recall aspects can be easier than some recognition aspects. For example, Pellicer-Sánchez and Schmitt (2010) found that the two recall components measured (word class and meaning) were learnt after the two recognition components (meaning and spelling); however, only the meaning component was tested at these two levels of sensitivity. Conversely, Pigada and Schmitt (2006) found that a recall component (spelling) was easier than some recognition aspects (grammatical knowledge). The scale in Figure 7 illustrates that there is a neat split between recognition and recall mastery in general,

with the recognition knowledge of all aspects being easier than any type of recall knowledge. If recall knowledge is more difficult than recognition knowledge, thus it will be expected to be acquired later. Therefore, the key finding is that recognition knowledge of *all* components was mastered (at least in terms of the criteria in our study) before *any* type of recall knowledge. In other words, as much previous research has suggested, recall mastery is hard for L2 learners (Pellicer-Sánchez & Schmitt, 2010; Read, 2000; Schmitt, 2014; Van Zeeland & Schmitt, 2013), and would appear after all the forms of recognition knowledge are mastered.

While recognition knowledge has been widely acknowledged as being greater and preceding recall knowledge for L2 learners (e.g., Laufer & Paribakht, 1998; Vermeer, 2001; Webb, 2008), this has usually been considered for only individual components. For example, Laufer and Goldstein (2004) demonstrated that recognition knowledge of the form–meaning link was easier than recall knowledge, forming the following order (from easier to more difficult) meaning recognition > form recognition > meaning recall > form recall. Only a few studies measured recognition and recall gains in various aspects of word knowledge (Schmitt & Meara, 1997; Webb, 2007a). They found that recognition knowledge was, in general, greater than recall knowledge, although they could not find a stable pattern between gains in recall and gains in recognition knowledge. Thus, these studies could not infer anything about the relationship between recognition and recall masteries across multiple aspects. My implicational scale involving recall and recognition knowledge of four different components builds on this research by attesting that the ease of recognition before recall knowledge applies to all of the components I measured. This supports the assumption that recall knowledge is more advanced than recognition knowledge (Read, 2000). Some teachers and learners might assume that if they can read a word, they should also be able to write it when needed. But reaching the point of full recall mastery (i.e., being able to think of and produce a word that is fully appropriate in terms of derivative form, collocation, and situation-based meaning) is extremely challenging. Indeed, it may take much longer to move from recognition to recall mastery than it takes to move from zero knowledge to recognition mastery. This IS suggests that the most important distinction in vocabulary knowledge might not be between the word knowledge components themselves (and which are more important or are learned before others), but rather between the recognition and recall mastery of those components.

It is important to mention that the IS does not mean there is absolutely no knowledge of a more difficult aspect (recall knowledge) until an easier one (recognition knowledge) is acquired. Indeed, based on the strong intercorrelations between aspects, it seems intuitive that some parallel learning of components is taking place until one particular aspect is mastered at the 80% or 75% level. In fact, Webb (2008) has claimed that in the process of learning the different recall and recognition aspects there is some parallel knowledge, so that partial recall knowledge might exist for most words learnt in recognition. The IS represents mastery of knowledge at the 80% accuracy level. Following Webb's results, some degree of recall knowledge might be present until the recognition knowledge is mastered at the 75-80% level. For example, learners may have partial/incomplete knowledge of the recall form–meaning link of a word whilst they are achieving mastery of collocational recognition knowledge. Overall, as it has been previously claimed (Henriksen, 1999; Melka, 1997; Nation, 2013; Read, 2000), this scale indicates that considerable recognition knowledge is needed before any degree of recall knowledge can develop.

4.5.3.2. Order of acquisition of individual components

The IS suggests that there is an ordering in the acquisition of the various vocabulary components. According to this order, the form–meaning link was the best known, and generally appears before other components, as shown in some previous studies (Henriksen, 1999; Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010; Tannenbaum et al., 2006). This indicates that the form–meaning link is likely to be the first aspect to be mastered, whilst the rest build on it (e.g., Agustín Llach & Moreno Espinosa, 2014; Schmitt, 1998, 2000; Schmitt & McCarthy, 1997). However, this result contradicts other previous multicomponent research (Chen & Truscott, 2010; Webb, 2005, 2007a, 2007b, 2009) which suggests that other components precede the acquisition of the form–meaning link. In Webb's studies, there was a tendency for knowledge of spelling and sometimes word class to develop before the form–meaning link, although this order varied according to number of encounters, treatment and/or the time on task. I did not measure spelling or word class knowledge, so I cannot comment directly on how they might fit into my acquisition order, other than to say that it is possible that some untested word knowledge components may prove to be easier than

the form–meaning link, i.e. the form–meaning link is not necessarily the beginning of a complete acquisition scale.

Chen and Truscott (2010) found that the form–meaning link was the least-known aspect for very low-frequency words. I also found that form recall of the form–meaning link was difficult for my learners for low-frequency words. In fact, this partially explains one of the surprising results of my study: that form recall of the form–meaning link proved more difficult than collocate form recall. Although this result may be counter intuitive, it can be partially explained by frequency. For the form–meaning recall test, students had to recall the form of words (the most difficult stage in the form–meaning link (Laufer, 1998; Laufer & Goldstein, 2004)) from a range of frequency levels (from 1-9K). For higher-frequency words, form–meaning recall was generally better known than collocation recall, but this was not the case for lower-frequency words (e.g., *grate* or *indent*), where learners struggled in recalling the written form. They found it easier to recall a collocate for the low-frequency target words once they were provided in the collocation recall test, especially as the collocates were always high-frequency words (within the first 3K frequency bands) and presumably known. Thus, frequency has an effect on the difficulty of the form–meaning recall test: learners can recall very frequent collocates of words slightly better than recalling the spelling of lower-frequency target words. Therefore, my results confirm the conclusion of several scholars (e.g., Barcroft, 2002; Laufer, 1998) that form recall is difficult, especially for lower-frequency words.

Knowledge of form concerning the syntactically constrained knowledge of word family members (as in the derivatives measures) also seems difficult and is learned relatively late (Barcroft, 2002; Chui, 2006; Jiang, 2000; Nagy et al., 1993; Schmitt & Meara, 1997). Indeed, it is not surprising that derivatives (both recognition and recall) are among the last aspects to be mastered. Yet, the development of derivational and morphological knowledge is of great importance for vocabulary growth (Kieffer & Lesaux, 2008, 2012b; Stahl & Nagy, 2006; Zhang & Koda, 2011). This finding, thus, not only corroborates arguments stating that form knowledge is one of the most difficult components to learn, but also that it probably requires explicit teaching attention to be acquired (Barcroft, 2002). My results also inform the ongoing debate regarding the best unit of measure for vocabulary size. Nation (2013) suggests the word family as the most useful unit, but this assumes that if a learner knows one word family member, they

would also know all the others. There is growing evidence that this is not the case, as learner knowledge of a family's derivatives cannot be assumed in either recall or recognition (e.g., Mochizuki & Aizawa, 2000; Saigh, 2015; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002; Ward & Chuenjundaeng, 2009). My results support this finding, and strongly suggest that the word family approach is probably too generous, over-estimating the derivative knowledge of learners, and thus should be reconsidered as the main unit of measure of vocabulary size. This adds strength to the argument that the lemma may be a safer unit to use in pedagogical studies (e.g., Brown, 2013; Kremmel, 2016; Kremmel & Schmitt, 2016; Treffers-Daller, Parslow, & Williams, 2016). Because the members of lemmas are more similar morphologically and semantically than members of word families, they can be expected to be recognised and known by students as the same unit. In fact, it has already been shown that the use of lemmas over word families shows advantages on certain vocabulary measurements (Treffers-Daller et al., 2016).

Collocations have traditionally been considered a difficult component, with many studies showing them as problematic for L2 learners (e.g., Chui, 2006; Peters, 2016; Shimamoto, 2000; Webb et al., 2013). Based on this previous research, it might be expected that they would be further towards the difficult end of the scale. My findings, however, show that when we measure collocations along with other aspects such as derivatives and multiple meanings, collocations are not the aspect which poses the most problems for L2 learners. I interpret this finding not as showing that collocations are easier than believed before, but as showing that derivative and multiple-meanings knowledge are harder than previously thought. Although there are some claims suggesting that recognition knowledge of collocations is known almost at the same level as recall knowledge of derivatives (i.e., Chui, 2006), this better knowledge of recall derivatives is due to the test requiring the production of only one derivative, as opposed to the four required in my study. Therefore, when knowledge of all the derivatives of a word is required, this aspect becomes one of the most difficult aspects to master, proving even more difficult than collocations.

The scale also shows that in the development of vocabulary, knowledge of multiple meanings is difficult and late in coming, both in the recognition and recall spectrum. Previous research suggests that the acquisition of even a single meaning is not straightforward. Van Zeeland and Schmitt (2013) reported that meaning recall was

the last aspect to be acquired incidentally from listening, after recognition of spelling and word parts. Pigada and Schmitt (2006) found that knowledge of meaning was acquired after spelling and knowledge of grammatical characteristics. Shimamoto (2000) discovered that meaning recall was more difficult than recognition of associations and recognition of collocations. Ma (2015) found that learning different meanings of a word is a further learning state which involves deeper vocabulary knowledge than learning the form. For these reasons, Wolter (2009) believes that meaning is a component learned late in the acquisition process. Although multicomponent research involving knowledge of multiple senses is scarce, there is some evidence of its difficulty. Schmitt (1998) looked at the acquisition of multiple meanings and found that L2 learners can achieve knowledge of other word knowledge components without having mastered a word's multiple meanings. It is not difficult to speculate why the multiple-meanings aspect is relatively problematic. Most pedagogical materials focus on single senses, and the amount of exposure required to come across less frequent meanings is exponentially greater than the most frequent sense. For example, in a 100-condordance-line random sample of the COCA Fiction corpus, the 'money' sense of *bank* occurred 68 times, the 'riverside' sense 19 times, the 'collection' sense (*bank of clouds*) 12 times, and the 'tilting' sense (*banking airplanes*) only once.

These scale results are likely also partially an artefact of the measures we used. For example, I asked for only a single collocation, but for four derivative forms. When only one derivative is required, collocations and derivatives seem to be similar in difficulty (e.g., Chui, 2006). Also, using tests requesting similar types of responses from the students as mine, Webb (2007a) found that after one, three and ten encounters of the word, the productive form–meaning link experienced lower gains, and was acquired after syntax knowledge (defined as "words which had a syntactic relationship beside each target word" (*ibid*.: 56), such as the relation between *letter* and *diary* to the target word *write*). Since this interpretation is similar to the description of collocations, his results suggest that aspects such as productive knowledge of collocations or grammatical functions (i.e., word class) may be acquired before meaning and form in the productive mastery. This highlights the fact that my scale is based on the particular methods of measurement employed, and that different methods may produce somewhat different results. Nevertheless, I feel the measures are all valid and reliable, and it is impossible to measure various aspects in exactly the same manner in any case.

Moreover, the validity of the scale is increased by the fact that it remains constant across the two different accuracy criteria used to consider an aspect achieved, 75% and 80%. This suggests that this order of acquisition of word knowledge components is robust and reliable, and not just an artefact of the scoring system. I therefore believe that my scale is the best description of the word knowledge acquisition order currently available, and only future research with different measures and learner populations will support/refute its generalisability. I suspect that although the acquisition order of components may change somewhat with different participants/measures, the recognition/recall distinction is likely to remain. The influence of the L1 factor will be studied in Chapter 7.

Although my scale is statistically robust and reliable, it could potentially be misleading in various ways. The scale indicates that the different word knowledge components follow a consistent difficulty order in their acquisition. However, this does not mean that they are strictly learned in sequence. The strong intercorrelations suggest that below the 75% criterion level, there may be some parallel learning accruing. For example, there is probably some degree of recall knowledge present before recognition knowledge is mastered at the 75-80% level. Also, individual words have their own idiosyncratic characteristics (e.g. cognateness, frequency) which can make some word knowledge components relatively easier or more difficult to acquire. Likewise, learners vary and may be better at picking up some components than others. Nevertheless, although the acquisition scale may not hold for any particular word or person, it represents group tendencies and is consistent across varying types of words, learning conditions, and learner proficiencies.

Finally, even though the aspects have different difficulties, the *degree* of difficulty between the implicational steps should not be seen as equidistant. The relative degree of difficulty was greatest between the collocation recognition and multiple-meanings recognition aspects, the multiple-meanings recognition and derivative recognition aspects, and the derivatives recall and multiple-meanings recall aspects. The difference between the remaining aspect pairs was less pronounced. Therefore, the time and effort it will take L2 learners in general to move from certain steps of the scale to others can be greater or lesser.

Taken together, the IS results demonstrate that, as it has been previously found for morphology studies (e.g., Ellis, 2008), the vocabulary knowledge aspects can be

expected to be learnt by L2 learners in a certain pattern. This finding confirms that, as compared to previous attempts to systematise vocabulary knowledge (e.g., Schmitt, 1997), more precise measures of word knowledge components and a larger sample were needed for a vocabulary sequence to be statistically supported. Moreover, it reinforces previous hypotheses that vocabulary knowledge aspects are not mastered in a 'know everything/know nothing' dichotomous way (Schmitt, 2010, 2014). Rather, vocabulary knowledge seems to be learnt in a sequential manner, although with knowledge of the different aspects developing in parallel along the continuum. According to these findings, recognition knowledge of all aspects would be mastered before any recall knowledge, with form–meaning link being the best-known aspect and derivatives and multiple-meanings the most difficult ones. Pedagogical implications of these findings will be discussed in Chapter 7.

4.5.4. Limitations and directions for further research

Although this study was carefully designed, there are some limitations that can become the starting point for future research. In the first place, the vocabulary knowledge construct is comprised of more aspects than practical constrains allowed me to include in this study. Thus, future research exploring additional components of word knowledge, such as constraints on use, grammatical functions and/or associations, is needed. Moreover, the present study was based only on reading and writing skills. Using factor analysis, Cheng and Matthews (2018) found that orthographic measures loaded onto a different factor from a phonological measure, indicating that written and oral aspects represent different, although related, constructs. This suggests that measuring oral ability might produce different findings to the ones presented in this study. Thus, it would be interesting to also test oral measures and compare how the relationships change when these aspects are included. Research combining the assessment of oral and written aspects would provide a more general and comprehensive picture of the overall vocabulary knowledge, solving one of the gaps of the present study.

Another limitation is the fact that this study examines the relationships of word knowledge components in English by speakers of only one L1. As a consequence, the conclusions are limited to Spanish learners of L2 English, and potentially to L1 speakers of Romance languages. It is, thus, still unclear whether these results are universal, and

therefore can be applied to any other EFL learners, or whether they are language-dependant. Therefore, replications of this study with other language populations would provide a more complete picture of the acquisition of vocabulary components in English as an L2.

Moreover, the scale presented in this study provides a cross-sectional analysis of the data. Research suggests that each word knowledge aspect has a changing nature. Thus, using only a cross-sectional study limits any conclusions about the evolution of the different word knowledge aspects, which has led some authors to propose longitudinal studies as the best way of understanding the progressive acquisition of word knowledge (Fitzpatrick, 2012). Therefore, further research employing a longitudinal study design in which we could compare not only the pattern of acquisition of word knowledge components at a particular point in time, but also the development that occurred between two times, would shed more light on these results. If it was found that the pattern remains, this would support the acquisition order of the vocabulary knowledge components, and thus it would have important implications for the area of teaching and testing vocabulary knowledge.

Furthermore, a valid question is whether changing the target word set would affect the results. In this sense, there are different possibilities that future research can examine, such as following the same target words' selection criteria but changing the specific set of target words, or changing the target words so that each recognition-recall combo aspect is tested with a different set of items. If the results remained largely the same after testing these other options, this would provide further evidence for the findings presented here. Additionally, as setting appropriate accuracy criteria to consider an aspect mastered is not something we can be completely confident about, future research could explore cut-off points other than 75 and 80%.

Finally, as in any research, the results are influenced by the way features are measured. Although our measurement instruments were carefully designed, there are many ways in which the different vocabulary aspects can be tested. I chose the specific tests for each aspects because I considered them the most appropriate for the type of knowledge, length and context of the study. However, with somewhat different measures, dissimilar results might be found, and it would be very interesting to explore how the findings vary with a different type of measurement.

4.6. Conclusion

Vocabulary knowledge comprises the acquisition of various components that, although interrelated, are different from each other and reflect various types of information about words. It has been claimed that acquisition of these different components is necessary to be able to use a word accurately and appropriately in different situations. However, few students seem to master multiple of those aspects, and it is still unknown how they are acquired. This study examined the knowledge that Spanish second language learners of English had of four aspects of word knowledge in recognition and recall: form–meaning link, multiple meanings, collocation and derivatives. The aim was to try and establish an order of difficulty of these aspects that could inform how they are learnt. Implicational scaling analysis was performed in order to reveal any acquisition orders of the different word knowledge components. The results show that it is indeed possible to suggest a pattern of difficulty, and thus an order of acquisition, of vocabulary knowledge aspects as a whole, despite the individual differences of words. In this pattern, recognition of the form–meaning link is considered the easiest aspect, followed by recognition knowledge of collocations, recognition knowledge of multiple meanings and recognition knowledge of derivatives. Once the recognition mastery of the different aspects are acquired, mastery of recall knowledge develops from recall knowledge of collocations, recall of the form–meaning link, derivative recall and multiple meanings recall. Therefore, it seems that that recognition mastery of a difficult aspect, such as derivatives, is acquired before recall mastery of a generally considered easier aspect, such as form–meaning link.

Chapter 5

Factors influencing knowledge of vocabulary components: Frequency, cognateness and proficiency.

As evidenced by Study 1 in Chapter 4, the various lexical components involved in knowing a word present different levels of difficulty for the learners. The amount of effort required to learn the multiple senses of a word, for example, is greater than for learning its form–meaning mapping. But the learning burden of a word not only depends on the specific type of knowledge of the word we are tackling. It is also affected by other intrinsic properties of that specific word, as well as general and metalinguistic knowledge of other languages. Whilst the effect of these factors has been reliably demonstrated for some aspects, there are still questions to be addressed. For example, what is the influence of these features on other word knowledge components than the form–meaning link, and the extent of their influence? If the different word knowledge components pose different learning burdens to students, the effect on learning difficulty of such factors may also differ across components. The present chapter reports on the extent to which factors that have been found to interfere in vocabulary learning affect the knowledge of different word knowledge components across the same set of lexical items. This will allow us to obtain a clearer picture of the influence of such factors on the order of acquisition reported in Study1 (Chapter 4).

5.1. Background of the study

5.1.1. What makes a word difficult? Factors affecting vocabulary acquisition

Learning vocabulary in a second language is a complex process. Learners not only need to acquire a large number of words to function in a L2 (e.g., Laufer &

Ravenhorst-Kalovski, 2010; Nation, 2006; Van Zeeland & Schmitt, 2012; Webb & Rodgers, 2009); they are also required to learn a lot of deeper information about those words if they are to understand and use vocabulary effectively in varied contexts (Kieffer & Lesaux, 2012a; Tannenbaum et al., 2006; Webb, 2005). On top of that, research has shown that there are a series of factors that affect the amount of effort required to learn specific words. These factors influencing the learning burden of vocabulary items can be both intralexical and interlexical, and have been found to affect different words and learners from diverse language backgrounds differently (Nation, 2013).

Intralexical factors are those which arise from the particular intrinsic features of words. Among the intralexical factors which have been found to increase word difficulty are: pronounceability, spelling, word length, morphological complexity, orthographic and phonological neighbourhood density, semantic-related features (i.e., abstractness, polysemy, register), conceptual difficulty, and word frequency (Laufer, 1997; Nagy, McClure, & Mir, 1995). Interlexical factors are related to the links between words in the target language and other languages, particularly the mother tongue. The L1 affects L2 learning in both positive and negative ways, and it has been found that L2 learners can acquire vocabulary by developing links between the two languages (Swan, 1997). The most widely cited and studied interlexical factors of word difficulty are language distance and cognate status.

The next section will focus on discussing some intra and interlexical factors that are of particular interest for the purpose of this study.

5.1.1.1. Intralexical factors: Word frequency and word length

The frequency in which words occur in a language affect the acquisition, processing, knowledge and use of vocabulary. For this reason, frequency is often considered and used as one of the best predictors of lexical knowledge (Nation, 2013; Schmitt, 2010). The logic behind it is that frequent words are the most likely to be encountered in discourse, and thus, more likely to be understood, processed and produced by learners more often (Ellis, 2002).

The effect of frequency on vocabulary acquisition has been widely acknowledged, with more frequent vocabulary being generally acquired earlier and

known better than less frequent words (e.g., Schmitt et al., 2001; Vidal, 2003, 2011). For example, Milton (2009) found that a word's frequency correlates with the likelihood of that word being known, especially for the most frequent items (i.e., first 4,000 words). Vidal (2011) studied how various factors, including corpus frequency, affected the incidental acquisition of vocabulary through reading and listening. She found that corpus frequency influenced vocabulary acquisition, with lower-frequency words being more difficult to acquire. However, the frequency contribution to vocabulary learning, although explaining 20% of the variance, was less than that of other factors. Thus, because a word's frequency has been found to be related to the likelihood of L2 learners knowing that word, most research includes this as a factor that predicts vocabulary knowledge or as a basis for the selection of words to be tested/taught.

This frequency effect has been found not only for single-word items, but also for multi-word units, showing that it accounts for a large part of the variance in vocabulary acquisition (e.g., Garnier, 2016; González Fernández & Schmitt, 2015; Redwood & Schmitt, 2011; Wolter & Gyllstad, 2013; Wolter & Yamashita, 2017). For example, González Fernández and Schmitt (2015) found that frequency was the best predictor of recall collocational knowledge (compared to *MI* and *t*-score) accounting for 20% of the variance. Similarly, Redwood and Schmitt (2011) showed that frequency predicted knowledge of phrasal verbs substantially, accounting for approximately 20% of the variance of recall knowledge of phrasal verbs and 9% of recognition knowledge. In a recent study, Wolter and Yamashita (2017) found an effect of both word-level and collocation-level non-lemmatised frequency on the collocational knowledge of native and non-native English speakers, with a less pronounced effect on intermediate students. Moderate to strong correlations have also been found between knowledge of collocations and frequency. In a Meta-analysis of nineteen published collocation tests, Durrant (2014) found that corpus frequency (BNC and COCA) correlates moderately with L2 learners' collocational knowledge, although this strength varies depending on the corpora used (i.e., COCA frequencies were more strongly related to learners' knowledge). Thus, these studies suggest that L2 learners seem sensitive to frequency of lexical units.

However, although it is clear that frequency influences lexical knowledge, its unique contribution does not appear to be overly large. Considering that frequent items are prevalent, it may seem surprising that the relationship between frequency and

vocabulary knowledge is not stronger. It has been found that some frequent words are difficult for learners, whilst some rare ones do not seem to pose a lot of trouble. In fact, some research has found inconsistent and even insignificant results regarding the effect of frequency on vocabulary (e.g., Macis & Schmitt, 2017a, 2017b; Pellicer-Sánchez, 2017; Szudarski & Carter, 2016). Part of the reason for this situation has to do with limitations of the word counts. The frequency counts of words are generally retrieved from corpora, and these counts may create problems for the prediction of knowledge of some words. Firstly, corpus data in general does not represent the language that a L2 learner has encountered, but rather particular types of discourse (Durrant, 2014; Durrant & Doherty, 2010). Secondly, the frequency counts in current corpora lack fine-grained semantic tagging, and thus frequency counts do not distinguish between the different senses of polysemous words. Occurrence counts for such words cannot accurately reflect the exposure a learner has to the word when it is used with a specific meaning. Thus, including the counts of all senses in the final frequency count of a word may lead to an overestimation of learners' knowledge based on their real experience. For these reasons, frequency counts derived from corpora may not be representative of knowledge of certain words (e.g., technical vocabulary), or certain word knowledge aspects. For example, since a polysemous word's frequency includes the occurrences of its multiple meanings (or some of them), this frequency may not be a significant predictor of specific meanings of that word. Similarly, the lemma frequency of a word may not have an effect on the knowledge of a word's derivatives, since it only includes the counts for its inflections.

Altogether, frequency is considered the best documented predictor of learner vocabulary knowledge to date (Diependaele et al., 2013; Nation, 2013) and the basis for the usage-based theories of language processing and acquisition. However, learning a language is too complex a process, and thus it can never be predicted by a single factor (Dörnyei, 2007; Ellis & Larsen-Freeman, 2006). Thus, lexical knowledge is likely to also be influenced by other features of a word.

Frequency has been claimed to be very much related to other intralexical factors such as word length, with more frequent words generally being shorter than less frequent ones (Milton, 2009). However, whilst the effect of frequency on vocabulary acquisition has been widely demonstrated, there are mixed findings regarding the effect of word length on lexical learning. It appears obvious to believe that longer words would

be more demanding for learners than shorter ones, and this has actually been demonstrated by previous studies (e.g., Barclay, 2017; Campoy, 2008; Peters, 2016). Campoy (2008) investigated the effect of word length on a form recognition learning task (both aural and visual) and found that shorter words were recognised better than longer words across both modes at the form–meaning link level. Barclay (2017) studied how word length affected acquisition and retention of words. He found that fewer longer words were learnt in both the form recall and form recognition masteries from visual input, and those that were learnt needed more exposures to be acquired. His results suggest that longer words are more difficult to acquire in both the recognition and production of form–meaning, but that word length did not affect retention. Peters (2016) studied how word length affected the acquisition of collocations from TV, and found that longer words presented a greater burden to L2 learners at the form–meaning recall mastery. However, Peters found the opposite effect at the form–meaning recognition level, with longer words appearing to be easier to remember. These mixed findings, along with the lack of evidence on how word length influences other word knowledge components than the form–meaning link, indicate it is interesting to evaluate its effect.

Overall, the inconclusive findings regarding the influence of frequency and word length on vocabulary knowledge require further investigation. Additionally, these effects have been studied mainly for the form–meaning mapping aspect. Consequently, it is still uncertain whether these factors affect different components of word knowledge and to what extent.

5.1.1.2. Interlexical factors: Cognate status and L2 proficiency

The influence of interlexical factors on vocabulary acquisition has also been explored, especially with regard to cognate status and proficiency or vocabulary size in the L2. Vocabulary proficiency level has been demonstrated to impact all aspects of language and lexical acquisition (Milton, 2013; Stæhr, 2008). The message is clear: learners with a higher vocabulary proficiency acquire more new words and better. Much previous research has shown that it is a factor in the acquisition of not only single words (Montero Perez, Peters, Clarebout, & Desmet, 2014; Zahar, Cobb, & Spada, 2001) but also formulaic language (Peters, 2016; Vilkaitė & Schmitt, 2017), and that not only it influences the acquisition, but also the retention of words (Vidal, 2011). Proficiency has

been found to correlate with other interlexical factors, such as cognate status (Malabonga, Kenyon, Carlo, August, & Louguit, 2008). Cognateness is a factor of particular interest in the present study.

Cognates are words which share similar pronunciation, spelling and meaning in two languages because they have a common origin. Cognates can share different degrees of phonological and/or orthographic form, and are always related semantically although they are not necessarily translation equivalents (Hall, 2002). Different types of cognates have been identified: i) phonologically similar and orthographically identical words (e.g., *dental – dental*), ii) phonologically similar but orthographically different words (e.g., *velocity – velocidad*), and iii) false cognates in which words are phonologically and orthographically similar but not related in meaning (e.g., *éxito* [meaning 'success']– *exit*) (Rodríguez, 2001). The specific characteristics of cognates mean that their learning process is likely to differ from that of other words. When a student is learning a language that shares cognates with their mother tongue, the linguistic commonalities are believed to provide learners with a "fund of knowledge" (Moll, Amanti, Neff, & González, 1992) that facilitates the acquisition of the L2 vocabulary. That is, learners may be able to draw on their lexical knowledge in the L1 to ease the learning of the L2. This effect is known as the "cognate facilitation effect", and has been described as the "advantage cognate words have over non-cognates in speed of recognition and production of words during language tasks" (Rosselli et al., 2012, p. 650). Research has found evidence of this effect in different pairs of cognate languages, such as Spanish-English (Bravo, Hiebert, & Pearson, 2005; Méndez Pérez et al., 2010), Italian-English (Siyambalapitiya, Chenery, & Copland, 2009) or Dutch-English (Van Hell & De Groot, 1998). The specific focus of this study will concern the Spanish-English pair.

Spanish and English are related languages that share a substantial number of cognate pairs (10,000-15,000, according to Nash, 1997). There is substantiated evidence that knowledge of Spanish-English cognates improves the knowledge that L1 Spanish learners have of L2 English vocabulary (Chen et al., 2012; Dressler et al., 2011; Hancin-Bhatt & Nagy, 1994; Nagy, García, Durgunoğlu, & Hancin-Bhatt, 1993; Proctor & Mo, 2009; Ramírez et al., 2013; Rosselli et al., 2012; Vidal, 2011). For example, Vidal (2011) studied how different factors affected the incidental vocabulary acquisition through reading and listening of 230 Spanish speakers learners of L2

English. She found that predictability from word forms and parts (i.e., cognates and morphologically transparent words (*short – shortly*) was the best predictor of vocabulary acquisition in both reading and listening, explaining 40-57% of the variance respectively. She concludes that cognates seems to be the best predictor of incidental vocabulary acquisition, and that L2 students "appear to attend mainly to those words that are similar to words in their L1" (Vidal, 2011, p. 241). Rosselli et al. (2012) studied how cognate status affected the form-recall ability of 117 Spanish-English balanced and unbalanced bilinguals. They found a high cognate effect, with cognates being significantly better known than non-cognates in all groups and independently of the target language. However, this effect interacted with the level of bilingualism, with unbalanced bilinguals knowing more non-cognates in the preferred than the non-preferred language and the balanced group producing cognates in similar proportions. Chen et al. (2012) examined the knowledge of English-Spanish cognates and non-cognates in English monolingual children, L1 Spanish, and L1 Chinese children learners of L2 English. They found that Spanish learners outperformed Chinese learners in knowledge of cognates, but were similar in their understanding of non-cognates. This finding suggest that cognate facilitation can lead to higher overall L2 vocabulary knowledge. Malabonga et al. (2008) designed a test to measure cognate awareness in bilingual Spanish-English students of three consecutive grades. They found that their performance in the awareness test correlated with vocabulary proficiency in Spanish, and that cognate recognition increases as children progress academically.

This sensitivity to cognates has been found even in children as early as kindergarten. Méndez Pérez et al. (2010) tested whether 89 bilingual children in kindergarten and first grade were able to recognise cognates of Spanish words, and how this ability affected their vocabulary proficiency in English. They found that cognates were produced more accurately than non-cognates, and that this knowledge of cognate words depended on the amount of exposure children had to both languages (i.e., proficiency). Children that were more exposed to Spanish knew more English cognates of Spanish words than those more exposed to English or with a balanced exposure to both languages. Simpson Baird, Palacios, and Kibler (2016) examined the knowledge of cognates and false cognates in emergent Spanish-English bilingual children (mean age 4.6 years old). Employing two English and Spanish tasks involving receptive and productive form–meaning link knowledge, they found a bilingual cognate advantage,

with cognates and false cognates being correctly identified and recalled in both languages.

Altogether, this increasing number of studies have shown that cognates facilitates the acquisition of L2 English vocabulary in L1 Spanish learners. However, it has also been claimed that even though cognates represent a potential linguistic advantage for vocabulary acquisition, bilingual students are generally unaware of the existence of cognates, and thus often fail to notice even transparent cognates as such (August et al., 2005; García, 1991; Lubliner & Hiebert, 2011; Nagy, García, et al., 1993; Otwinowska-Kasztelanic, 2009). For example, Nagy, García, et al. (1993) found that L1 Spanish-L2 English bilingual children in fifth and sixth grade selected less than half of the known cognate words that they found on a cognate identification test. They also found that knowledge of the Spanish cognates was related to their knowledge of the meaning of those English words, but only if learners were able to identify that those words are cognates. Similarly, García (1991) documented that bilingual readers often showed no awareness of even cognate pairs that have parallel forms in Spanish and English. These results suggest that cognate awareness is necessary for the facilitation effect and transfer of L1 lexical knowledge to L2 vocabulary acquisition to occur.

Since the potential facilitation effect of cognates depends on cognates being noticed and recognised, it is important to identify which features facilitate students' awareness of cognates and cross-language transfer. Different factors have been claimed to affect cognate awareness, such as degree of bilingualism, L1 and L2 proficiency, metalinguistic knowledge, frequency of the word, or orthographic and phonetic similarity (Chen et al., 2012; Dressler et al., 2011; Malabonga et al., 2008; Nagy, García, et al., 1993). The degree of orthographic similarity is a commonly cited feature affecting learners' recognition of cognates. For example, Nagy, García, et al. (1993) discovered that students were more successful at noticing and recognising cognate pairs when their forms clearly overlapped (e.g., *animal* - *animal*), and that even minor spelling differences significantly reduced their ability to identify cognates. However, even orthographic transparency is not enough to cross-transfer cognate words from the L1 to the L2. The semantic commonality seems to be the golden factor that determines this successful cross-linguistic transfer. In the case of Spanish and English, the different evolution of each language means that two cognate pairs with similar form may not always have precisely similar meanings today. This is the case of false cognates (e.g.,

molestar [meaning 'disturb'] and *molest* [meaning 'abuse sexually']), which can in fact hinder the acquisition of an L2 word (Lubliner & Hiebert, 2011). These findings indicate that the more overlap between these dimensions (form and meaning), the easier the cross-linguistic transfer of those cognates, whilst the greater the differences, the more difficult to recognise cognates, and thus the less the facilitation effect (Lubliner & Hiebert, 2011; Schwartz, Kroll, & Diaz, 2007).

An interesting feature of Spanish-English cognate pairs is that whilst the words in Spanish are generally high-frequency words, they are low-frequency words in English. Lubliner and Hiebert (2011) found that in a corpus of general academic words (including the GSL and AWL) 77% of the cognates were more frequent in Spanish than English (e.g., *acquire* and *adquirir*). This fact has led some researchers to propose the frequency of the cognate word as another potential factor that influences the effect of cognates on vocabulary acquisition (Chen et al., 2012). They suggest that it is possible that cognate awareness is particularly useful for learning low-frequency English words by L1 Spanish learners, as most low-frequency English words are high-frequency Spanish words. This claim suggests that frequency and cognateness may interact in their effect on vocabulary knowledge.

Although the effect of these interlexical and intralexical factors on vocabulary knowledge has been substantially investigated, the object of this previous research has primarily been the form–meaning link aspect of words. Consequently, little is known about whether these factors influence other word knowledge components, such as derivatives or multiple meanings. This study aims to explore the extent to which frequency, cognateness, word length and proficiency interfere in the knowledge of different vocabulary components. Specifically, I am interested in whether cognateness is a better predictor of vocabulary knowledge than frequency, and, if so, for which aspects.

5.2. Purpose and research questions

After having examined the knowledge that Spanish EFL learners have of multiple word knowledge components and how they are acquired, I wanted to investigate the effect that different inter- and intralexical factors have on these various types of knowledge. The present study aims to gain further insight into the influence of

such factors on the form–meaning link aspect, whilst also understanding how the other word knowledge components are affected. More precisely, it compares the effects of factors that have been found to affect vocabulary learning, that is, cognateness, frequency, word length and proficiency, on the four recognition and recall word knowledge components across the same set of lexical items. Mixed-effects modelling was chosen to examine the effect of these lexical features on word knowledge components.

In this study, the following research questions are addressed:

1. Which inter- and intralexical factors affect the knowledge of different word knowledge components?
2. Do these factors influence the various word knowledge components equally?
3. Is there a single factor that can be identified as the strongest contributor of knowledge across all aspects?
4. Does the cognate status of certain items affect the knowledge of all the vocabulary knowledge components or only the form–meaning link?
5. Which factor, frequency or cognateness, is a better predictor of vocabulary knowledge?

5.3. Methodology

The measurement instrument of this study is that described in Chapter 3, and the participants those presented in Chapter 4. The 20 target words in the measures were divided into cognates (n = 7) and non-cognates (n = 13). The cognates were selected based on my intuitions as a Spanish native speaker and their most frequent meaning double-checked in the Corpus del Español del Siglo XXI (CORPES XXI) (Real Academia, 2015), the most comprehensive corpus of Spanish comprised of 225 million words from all varieties of Spanish language. Two Spanish native speakers teachers of EFL English were consulted on the status of each target word, and there was 100% agreement in the target words considered cognates. Both cognates and non-cognates were matched at various levels: 1) their range of frequency bands (1K-6K and 1K-9K respectively) (cognates M = 2.57, SD = 1.90, non-cognates M = 3, SD = 2.68, t(18) = -

.374, p =.71), 2) their word length in letters (cognates M = 7 letters, SD = 2.45, non-cognates M = 6.2 letters, SD = 1.95, $t(18)$ = -.85, p =.41), and 3) their range of word classes, including all word classes in both sets (cognates: 3 nouns (42.9%), 2 adjectives (28.6%), 2 verbs (28.6%); non-cognates: 4 nouns (30.8%), 3 adjectives (23.1%), 6 verbs (46.2%), $t(18)$ = -.174, p =.86). Participants had a variety of proficiency levels, as shown by the VLT scores' range (see Table 11). The results of the recognition and recall word knowledge tests by Spanish participants were analysed by means of Mixed-effects modelling analysis in order to address the research questions. Table 3 shows the target words' frequency, cognateness status and word length.

5.3.1. Mixed-effects modelling analysis

In order to explore whether the inter- and intralexical factors presented above affect knowledge of the different word knowledge components, I computed generalised linear mixed-effects model analyses. Mixed-effects modelling was chosen because of its advantages over traditional types of regression analysis (i.e., ANOVA). First, it allows for the inclusion of both subject and items as random effects in addition to fixed effects in the same model, eliminating the need for separate by-subject and by-item analyses (Quené & Van den Bergh, 2008). This also allows me to account for individual variation by-subject and by-item, as well as for certain experimental effects such as learning or fatigue (Baayen, Davidson, & Bates, 2008; Winter, 2013). Also, mixed-effects models do not assume homogeneity of variance, which reduces the chances of model overfit and thus makes it more likely to be generalised to the entire population of learners and similar items (Agresti, 2002). Moreover, these models allow assessing effects and interactions of both continuous and discrete variables in the same model (Cunnings, 2012; Quené & Van den Bergh, 2008). Overall, the use of mixed-effects modelling analysis allows the identification of predictors with confidence.

The data was analysed using generalised linear mixed-effects models (aka mixed logit model (Jaeger, 2008)), as they are more suitable for the analysis of binomial and count outcomes. The models were fitted using the *lme4* package (Bates et al., 2014) version 1.1-13 in R-Studio Software (R Core Team, 2017), version 3.4.1. The final dataset comprised 2880 responses, from 144 participants to 20 words.

The main condition of interest for this analysis, cognateness status, was coded as a two-level categorical variable (cognates vs. non-cognates). The frequency variable represents the raw lemma frequency of each target word in number of occurrences in the whole COCA corpus. Gender and age were added to the analysis as covariates to control for their potential effect. Only the Lemma Frequency variable was log-transformed for further analysis, in order to reduce the skewedness of the distribution and because it was measured in a different scale than the remaining variables. The dependent variables (i.e., scores in each word knowledge test) were not transformed, as they include values from 0-4 and log transformation is impossible for zero values. Finally, Gender, Age, Proficiency and Word Length were used as raw values because the models did not experience any improvement when the log transformation was applied to them. Thus, since log transformation makes results more difficult to interpret, I decided to use the non-transformed data, with the exception of Frequency. Table 19 presents the range, mean, and the standard deviation of these variables.

Table 19 Range, mean and standard deviation of variables (Lemma Frequency before log-transformation).

Variable	Min	Max	Range	Mean	*SD*
Age	18	65	47	25.25	8.02
Lemma Frequency	136	336,890	336,754	62,303.9	80,726.5
Word Length	4	11	7	6.45	2.06
Proficiency	32	119	87	81.37	20.08

The predictor variables did not correlate significantly, and therefore they were all included as potential important fixed effects in the models. According to Barr, Levy, Scheepers, and Tily (2013), failure to include random effects in the model increases the chances of Type 1 error considerably, and thus, participants and items were included as random intercepts. Since the outcome variables were scored differently (0-1, 0-3 or 0-4), different types of regressions were employed for each model. When the dependent variables were binary (i.e., Form–meaning Recall, Form–meaning Recognition, Collocation Recall, and Collocation Recognition), I used generalised linear models with binomial regressions. When they were multinomial/count variables, (Derivative Recall,

Derivative Recognition, Multiple-Meanings Recognition and Multiple-Meanings Recall), I employed generalised linear models with poisson distribution.

The procedure of model development was conducted as follows. Eight separate models (one for each of the eight dependent variables) were developed. Model fitting always started from a core model, which included the independent variables frequency and cognateness status as fixed effects, the score of the particular word knowledge component as the dependent variable, and random effects for subject and item. These two main predictors of knowledge were retained in the models regardless of their significance, as their comparison was important in order to answer the research question about which of these two factors are better/stronger predictors of vocabulary knowledge.

All the potential explanatory variables were included in the model as predictors of knowledge: log lemma frequency, cognateness status, age, gender, proficiency and word length, with participants and items as random variables. Starting from this full model, when a variable was found not to statistically contribute to the fit of the model, I proceeded by employing a backwards step-by-step model selection procedure removing such non-significant variables (Manning, 2007). This process involved dropping one-by-one the non-significant variables starting from those with the lowest *z*-score and then refitting the model. This procedure continued until all insignificant variables were eliminated, except cognateness status and lemma frequency, which were considered the main predictors and thus retained in the model. Every time the model was refitted, it was explicitly compared to the previous one using the likelihood ratio test to check if dropping the variables did not significantly change the model fit. That is, if the likelihood ratio test yielded a significant *p* value[13] (Siyanova-Chanturia & Spina, 2015). The interactions between proficiency and cognateness, and frequency and cognateness were also examined, but only kept in the model if the fit was significantly better compared to the previous model. If an interaction showed to be statistically significant it was further analysed graphically (plotted using the *effects* package (Fox & Hong, 2009)). Significance values were estimated by the R package *lmerTest* (version 2.0–33; Kuznetsova, Brockhoff, & Christensen, 2016). The final best-fitting model is the one reported for each word knowledge component.

[13] Significance level for all the analysis was $p < .05$.

5.4. Results and interpretation

5.4.1. Preliminary analysis: Descriptive statistics and correlations

In order to obtain general information about whether the two main factors of interest in this study, frequency and cognateness, have some effect on the different components, some preliminary analyses were conducted.

First, correlational analysis was computed in order to find out whether the lemma frequency of each of 20 target words was related to the average score for each target word in the eight word knowledge aspects.

Table 20 Correlation between words' lemma frequency (COCA) and knowledge in the different aspects (Spanish group).

	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
Lemma Freq	.506*	-.108	.660**	.475*	-.099	.848**	.493*	.524*

Spearman: **p < .01, *p < .05

FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition

Table 20 shows that the Lemma Frequency of the 20 items correlated positively (and was statistically significant) with the total of correct responses for six of all the aspects except Collocation Recall (p = -.1) and Collocation Recognition (p = -.1). Given that in this analysis I am accounting for the frequency of the target word alone, it seems logical that only this frequency count does not have an effect on knowledge of the collocate word. This result suggests that the collocate frequency, as well as the frequency of the collocation as a whole, may be needed for the effect of frequency to be attested with collocational knowledge. Overall, the findings seem to indicate that frequency has some influence on the knowledge of most aspects.

As for the cognate status, in order to find out whether the participants' knowledge of cognate words was generally better than their knowledge of non-cognate

words for each component, the average percentage score[14] of correct answers to cognate words (n = 7) and to non-cognate words (n = 13) per participant in each of the eight word knowledge tests were calculated. Table 21 shows these descriptive statistics.

Table 21 Mean percentage score for cognateness status in all aspects.

Aspect	Cognates (%)	Non-Cognates (%)	Mean difference (%)
Form–meaning Recall	51.59 (21.04)	54.7 (20.81)	-3.11
Form–meaning Recognition	86.01 (12.7)	81.10 (17.14)	4.91
Derivative Recall	49.43 (20.31)	52.83 (21.78)	-3.4
Derivative Recognition	61.90 (18.81)	60.64 (21.24)	1.26
Multiple-Meanings Recall	58.73 (14.85)	44.82 (18.0)	13.91
Multiple-Meanings Recognition	76.13 (13.56)	67.88 (17.64)	8.25
Collocation Recall	64.39 (23.29)	58.65 (19.25)	5.74
Collocation Recognition	84.45 (17.87)	77.14 (19.19)	7.31

Note: standard deviations presented in parenthesis.

The mean scores suggest that cognate words are generally better known than non-cognate words in most apects, except Form–meaning Recall and Derivative Recall. The difference is usually quite small, and this result may indicate that there are other features, apart from cognateness, that affect knowledge of these components. For example, this slightly higher score in derivatives knowledge of non-cognates may indicate that, when encountering L2 words that are similar to words in their mother tongue, learners may be inclined to applying the derivational rules that the form in the L1 follows, instead of the particular suffixes of the L2 form. This practice would lead to more mistakes in the cognate words, whereas for the non-cognate words they only have the L2 suffixes to be applied and thus are not confused by L1 interference. However, the differences in knowledge for the derivatives and form–meaning are so small that they are probably best seen as not being meaningful. Mixed-effects modelling is needed in order to obtain a more comprehensive understanding on the effects of cognateness and frequency on the vocabulary aspects.

[14] The percentage scores were employed for this cognate analysis in order to make the scores of the different word knowledge components tests comparable.

5.4.2. Model analysis: Mixed-effects modelling

This section presents the models for each of the word knowledge components. For the Form–meaning link recall component the complete progression of model development is reported in order to illustrate the procedure employed. For subsequent aspects, although the complete procedure was conducted, only the final best-fitting model is presented, plus a table displaying a comparison of the fit indices to decide on the best-fitting model. The models for each vocabulary component are briefly discussed individually, leaving the more general discussion of findings for the Discussion section.

5.4.2.1. Form–Meaning Recall

Table 22 provides a summary of the results for the core model, which included lemma frequency and cognateness as predictors of Form–meaning Recall knowledge, along with random effects of subject and item[15].

Table 22 Core model of form–meaning recall knowledge.

	Form–Meaning link Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-8.328	1.270	-6.555	***
Cognateness status	0.090	0.517	0.174	.86
Lemma Frequency	0.866	0.122	7.096	***
Random Effects	***Variance***	***SD***		
Subject	1.81	1.34		
Item	1.16	1.1		

*** $p < .001$.

As Table 22 shows, cognateness status did not come out as a significant predictor in this model. Lemma frequency, on the other hand, was a strong and significant predictor. So, the results of the model show that high-frequency words were answered significantly better than low-frequency words in the Form–meaning Recall

[15] A negative β coefficient shows decreasing knowledge in the specific word knowledge component, while a positive β indicates increasing knowledge. Following the coding scheme used, the coefficient for cognateness status shows the direction of the non-cognate words. Thus, a positive value indicates better scores in non-cognates, and a negative value better scores in cognates.

aspect. Previous studies have reported a reliable effect of word frequency on L2 vocabulary knowledge, where higher-frequency words were recalled better (Vidal, 2011).

To check whether the rest of words' features (i.e., word length) and participants' characteristics (i.e., age, gender and proficiency) had an effect on the knowledge of the Form–meaning Recall link, a more comprehensive model including all the potential covariates as fixed effects was built. The results of the model showed that only proficiency was significant, along with lemma frequency (Table 23).

Table 23 Full model of form–meaning recall knowledge.

	Form–Meaning link Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-12.398	1.659	-7.473	***
Cognateness status	0.059	0.528	0.111	.91
Lemma Frequency	0.856	0.123	6.989	***
Age	-0.008	0.010	-0.774	.44
Gender	0.074	0.169	0.439	.66
Word Length	-0.049	0.122	-0.396	.69
Proficiency	0.057	0.004	13.551	***
Random Effects	***Variance***	***SD***		
Subject	0.46	0.68		
Item	1.16	1.1		

*** $p < .001$.

A backwards stepwise procedure was used to eliminate variables that did not significantly improve the overall fit of the model. The order of elimination of insignificant variables for the Form–meaning Recall model was the following: word length, gender, and age. Explicit model comparison confirmed that removal of these variables did not significantly deteriorate the overall model fit (i.e., the more complex model was not a significant improvement). At the end of this procedure, the best fitting model was the model with cognateness, frequency and proficiency as direct effects, that is, the core model plus proficiency (see Table 24).

Table 24 Model of form–meaning recall knowledge.

	Form–Meaning link Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-12.910	1.343	-9.607	***
Cognateness status	0.100	0.518	0.193	.85
Lemma Frequency	0.860	0.122	7.031	***
Proficiency	0.056	0.004	13.566	***
Random Effects	***Variance***	***SD***		
Subject	0.46	0.68		
Item	1.16	1.1		

*** $p < .001$.

This lack of a cognate effect on the form–meaning recall aspect is due to a lack of variation in the participants' responses to the cognate and non-cognate words. According to the mean scores in Table 21, the non-cognates were known almost at the same level as cognates, and this may have hindered this influence becoming significant. This result suggests that there might be another factor moderating the cognate effect at this level of knowledge.

The next step was to test whether there existed an interaction between the effects of the lemma frequency and cognateness status on the vocabulary knowledge. Research has claimed that a factor that affects cognate awareness is the frequency of the word (Chen et al., 2012; Dressler et al., 2011; Malabonga et al., 2008; Nagy, García, et al., 1993). Since it has been found that a high proportion of the Spanish-English cognates are more frequent in Spanish than English (e.g. *acquire* and *adquirir*) (Lubliner & Hiebert, 2011), an interaction between frequency and cognateness status may have been expected. Alternatively, it could be that frequent cognate words are known at the form–meaning recall level better than frequent non-cognates, and low-frequency cognates better than frequent non-cognates. A model including the interaction effect of frequency and cognateness plus any significant fixed effect (in this case only proficiency) was fitted. No evidence of such interaction was found for knowledge of Form–meaning Recall ($z = -.975$, $p = .33$), and explicit model comparison showed that inclusion of this interaction did not significantly improve the model fit (χ^2 (1) = .94, $p = .33$).

Finally, the interaction between the effects of cognateness and proficiency was tested, to see whether the knowledge of the recall form–meaning link of cognate and

non-cognate words improves with the proficiency of participants. This final model has three significant predictors (cognateness status, lemma frequency, and proficiency) and a significant interaction between cognateness and proficiency.

Table 25 Main effects on form–meaning recall knowledge.

	Form–Meaning link Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-11.962	1.361	-8.783	***
Cognateness status	-2.238	0.706	-3.169	**
Lemma Frequency	0.899	0.124	7.277	***
Proficiency	0.040	0.005	7.745	***
Cognateness*Proficiency	0.029	0.006	4.892	***
Random Effects	***Variance***	***SD***		
Subject	0.86	0.70		
Item	1.17	1.10		

*** $p < .001$, ** $p < .01$.

ANOVA comparison between the model with only direct effects of cognateness, frequency and proficiency, and the model adding the interaction between cognateness and proficiency showed that adding the interaction significantly improved the model fit, χ^2 (1) = 23.21, p <.001.

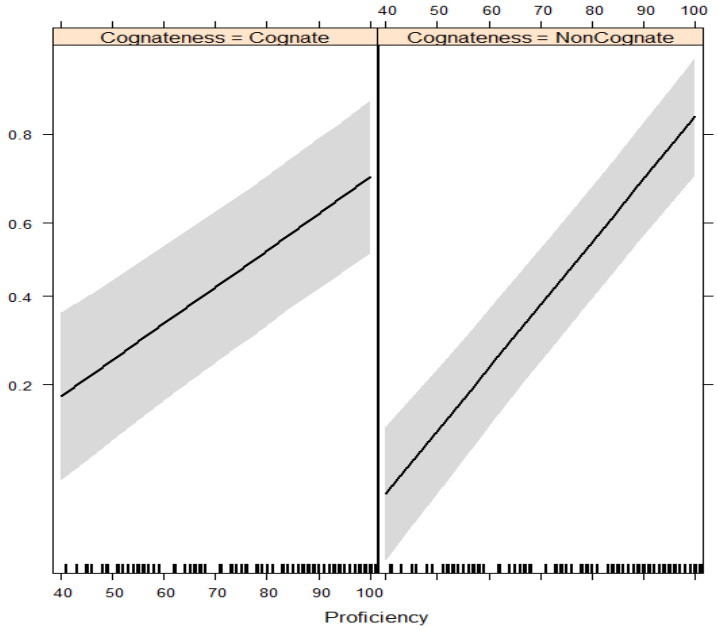
The indices used to decide on the best-fitting model for this analysis were the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). AIC and BIC are goodness-of-fit indices employed to compare the fit of models estimated with the same data, in order to decide which is more parsimonious (see also Section 6.2.4.3. in Chapter 6). The parsimony principle in statistics refers to the idea that the simplest explanation about the greatest number of observations is logical and favoured over more complex ones (Kline, 2016). Smaller values of the AIC and BIC indicate more parsimonious, and thus, a better fit, and so the model with the lowest AIC and BIC is the best-fitting model. According to these values, Model 1.3 (described in Table 25), proved to be the best-fitting model for the Form–meaning Recall aspect, since all the effects came out as significant (see Table 26). This model comprises only direct effects of cognateness, frequency and proficiency, and the interaction effect between cognateness and proficiency.

Table 26 Comparison of model fit for the form–meaning recall aspect.

	Models	AIC	BIC
1	Full model	2,482.4	2,536.1
1.1	minus word length	2,480.6	2,528.3
1.2	minus gender	2,478.8	2,520.5
1.3	minus age	2,477.3	2,513.1
1.3	Cognates* Lemma Frequency	2,478.4	2,520.1
1.3	Cognates*Proficiency	2,456.1	2,497.9

In this final model (Table 25), cognateness emerged as a significant direct predictor. In fact, it proved to be the strongest predictor of Form–meaning Recall knowledge, followed by frequency, proficiency, and the interaction between cognateness and proficiency. Since the cognateness variable takes the direction of non-cognates, the negative β value indicates that cognate words were significantly better known in the form–meaning recall aspect when controlling for word frequency, proficiency and the interaction between cognateness and proficiency. As cognates represent words with similar spelling and meaning in the L1 and L2, the significant and strong cognate effect on the recall form–meaning seems expected, as much previous research has reported (e.g., Rosselli et al., 2012). The interaction between cognateness and proficiency is plotted in Figure 8 using the *effects* package.

Figure 8 Interaction between cognateness and proficiency in form–meaning link recall.



This interaction suggests that the higher the proficiency of the participants, the higher the knowledge of both cognates and non-cognates (Figure 8). Very high-proficiency students seem to know the non-cognates better than the cognates, while very low-proficiency learners, as expected, knew the non-cognates worse than the cognates. It also seems that, as far as cognateness is concerned, the proficiency of the learners had a larger effect on the likelihood of knowing a word for non-cognates than cognates.

To sum up, the logistic regression analysis showed a clear effect of the lemma frequency, proficiency and cognateness on knowledge of the Form–meaning Recall link, as well as an interaction effect between the latter two. These findings are fully in line with previous research showing that the frequency and cognateness of words facilitate knowledge at this level. Age, gender or word length do not seem to be significant predictors of knowledge of the Form–meaning Recall component.

5.4.2.2. Form–Meaning Recognition

For brevity, for the remaining seven word knowledge tests, only the final best-fitting model is reported. Note though, that in each case, the complete procedure illustrated above was followed to develop that final model. Table 27 shows the comparative fit indices for the different fitted models, and thus the reason for the decision of the model with the cognateness and proficiency interaction as the best one.

Table 27 Comparison of model fit for the form–meaning recognition aspect.

	Models	**AIC**	**BIC**
1	Full model	1,792.9	1,846.6
1.1	minus gender	1,790.9	1,838.6
1.2	minus age	1,790.4	1,832.1
1.2	Cognates* Lemma Frequency	1,791.9	1,839.7
1.2	Cognates*Proficiency	1,784.2	1,831.9

Table 28 presents the best-fitting model of Form–meaning Recognition knowledge. Age and gender did not come out as significant predictors, and thus were not included in the model, with no significant deterioration of it. Similarly, no interaction effect was found between cognateness and lemma frequency (z = -.68, p

=.50), and explicit comparison of the model with and without the interaction showed no significant improvement of the model fit (χ^2 (1) = .45, p = .50).

Table 28 Model of form–meaning recognition knowledge.

	Form–Meaning link Recognition			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-8.246	1.867	-4.418	***
Cognateness status	-2.114	0.799	-2.644	**
Lemma Frequency	0.544	0.135	4.014	***
Word Length	0.395	0.145	2.722	**
Proficiency	0.040	0.006	6.318	***
Cognateness*Proficiency	0.022	0.007	2.966	**
Random Effects	***Variance***	***SD***		
Subject	0.23	0.48		
Item	1.45	1.21		

*** $p < .001$, ** $p < .01$.

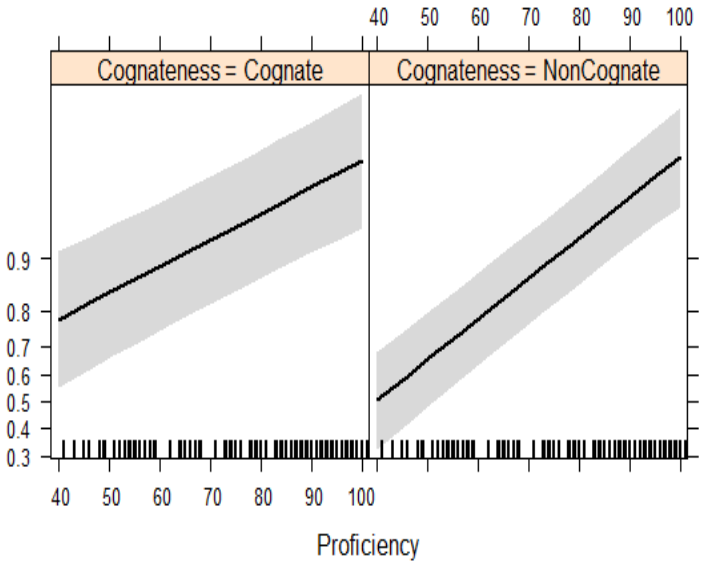
Thus, the final model included cognateness status, lemma frequency, proficiency and word length as significant fixed effects, and a significant interaction between cognateness and proficiency. ANOVA between the full model with proficiency and word length, and that same model with the proficiency and cognateness interaction showed that the model fit significantly improved by adding this interaction (χ^2 (1) = 8.22, p <.01). As in the case of form–meaning recall knowledge, cognateness status proved to be the strongest predictor of recognition of the form–meaning link. This supports previous research showing that the cognate word knowledge from the L1 Spanish transfers to recognition vocabulary knowledge in English (Méndez Pérez et al., 2010).

As expected, frequency and proficiency also contributed to the model, suggesting that high-frequency words were known better than low-frequency words and that learners with a higher proficiency perform better in the recognition knowledge of the form–meaning link. Word length came as a predictor of knowledge of the recognition form–meaning link, in line with previous research that shows that word length affects the recognition of words (e.g., Campoy, 2008). However, my results suggest that it was the longer words that were known better at the form–meaning

recognition level. These findings support recent results by Peters (2016) that long words are connected to better recognition than recall knowledge at the form–meaning level.

The interaction between cognateness and proficiency is plotted in Figure 9.

Figure 9 Interaction between cognateness and proficiency in form–meaning link recognition.



As in the recall form–meaning link, the interaction between cognateness and proficiency suggests there was a difference in the effect of proficiency on cognateness. Non-cognates seem to be slightly more affected by proficiency than cognates, and lower-proficiency learners seem to know cognates better than they knew non-cognates at the form–meaning recognition level. However, high-proficiency learners recognised the form–meaning link of cognate and non-cognate words similarly.

5.4.2.3. Derivative Recall

Table 29 shows the comparative fit indices for the different fitted models on which the decision of best-fitting model was based.

Table 29 Comparison of model fit for the derivative recall aspect.

	Models	AIC	BIC
1	Full model	8,510.0	8,563.7
1.1	minus gender	8,508.4	8,556.1
1.2	minus word length	8,508.2	8,549.9
1.2	Cognates* Lemma Frequency	8,510.1	8,557.8
1.2	Cognates*Proficiency	8,507.0	8,555.7

The final model is reported in Table 30. Gender and word length were not significant predictors, and thus are not included in the model. No evidence of an interaction effect between cognateness and frequency was found (z = .33, p = .74), with no significant improvement of model (χ^2 (1) = .10, p = .75). As opposed to the form–meaning link aspects, no significant interaction between cognateness and proficiency was found for the recall derivatives knowledge (z = .39, p = .70, no significant improvement of the model χ^2 (1) = .15, p = .70).

Table 30 Model of derivative recall knowledge.

	Derivative Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-1.281	0.184	-6.959	***
Cognateness status	0.065	0.067	0.963	.33
Lemma Frequency	0.043	0.015	2.777	**
Age	-0.006	0.002	-2.873	**
Proficiency	0.020	0.000	22.474	***
Random Effects	***Variance***	***SD***		
Subject	0.01	0.11		
Item	0.02	0.13		

*** p < .001, ** p < .01.

As Table 30 shows, the best fitting model included only direct effects of age, frequency, proficiency and cognateness. Lemma frequency came out as a significant predictor, indicating that higher-frequency words were known better than lower-frequency words. It was in fact the strongest predictor, suggesting that the lemma

frequency of words is the factor that influences knowledge of derivatives recall to the largest extent. Proficiency also proved to significantly contribute to recall knowledge of derivatives, with higher proficiency students performing better in this aspect than lower-proficiency students. Age also came out as a predictor. Previous research has found age to be an important factor in the acquisition of derivatives, with the development of derivative knowledge occurring late in the teenage (Nagy, Berninger, & Abbott, 2006). However, these studies have found that older age was associated with better performance in morphological tasks (e.g., Chondrogianni & Marinis, 2011), while my findings indicate that younger learners knew more derivatives in recall mastery than older learners. This surprising result may be an indication of younger adult learners having become more aware of derivatives. Finally, cognateness status did not come out as a significant predictor of recall derivative knowledge. This indicates that cognate words do not seem to present an advantage for the acquisition and knowledge of recall derivatives. One of the reasons for this result may be that cognate words do not form their derivatives in the exact same manner in the L1 and L2. This way, when L2 learners try to add to cognate words the derivative suffixes that apply to that word in their L1, they make mistakes.

5.4.2.4. Derivative Recognition

Table 31 shows the comparative fit indices for the different fitted models.

Table 31 Comparison of model fit for the derivative recognition aspect.

	Models	AIC	BIC
1	Full model	9,049.1	9,102.8
1.1	minus gender	9,047.1	9,094.9
1.2	minus age	9,046.8	9,082.6
1.3	minus word length	9,045.3	9,081.1
1.3	Cognates* Lemma Frequency	9,047.3	9,089.1
1.3	Cognates*Proficiency	9,043.0	9,081.7

Table 32 shows the final best-fitting model for derivative recognition knowledge. No significant direct effect of gender, word length or age was found, and thus they are absent from this model. No interaction effect between cognateness and

frequency was found ($z = 1.24$, $p = .22$), and explicit model comparison showed that the inclusion of this interaction did not significantly improve the model fit (χ^2 (1) = 1.47, $p = .23$). However, the cognateness and proficiency interaction was found to be significant, with ANOVA showing a significant improvement in the model fit after adding this interaction (χ^2 (1) = 5.82, $p < .05$). Therefore, the final best-fitting model included cognateness, lemma frequency and proficiency as significant direct predictors of recognition derivatives plus a significant interaction between cognateness and proficiency.

Table 32 Model of derivative recognition knowledge.

	Derivative Recognition			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-0.715	0.199	-3.594	***
Cognateness status	-0.302	0.135	-2.240	*
Lemma Frequency	0.054	0.017	3.213	**
Proficiency	0.013	0.001	11.458	***
Cognateness*Proficiency	0.003	0.001	2.421	*
Random Effects	***Variance***	***SD***		
Subject	0.01	0.09		
Item	0.02	0.14		

*** $p < .001$, ** $p < .01$, * $p < .05$.

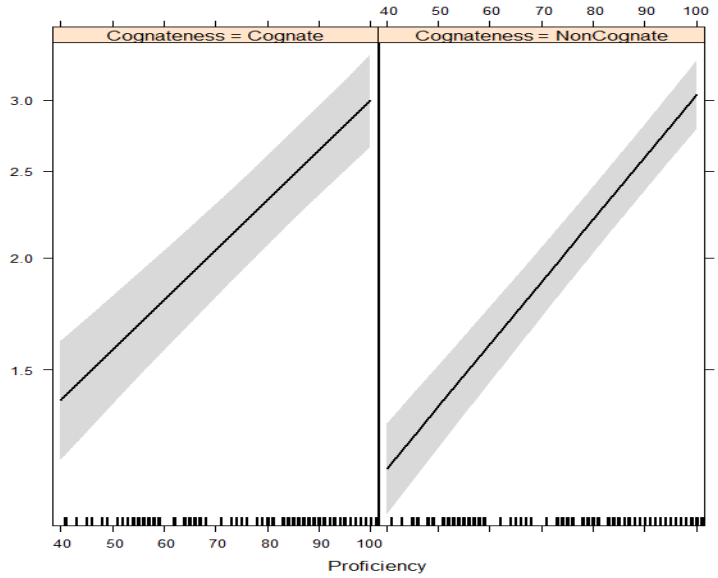
Cognateness status proved to be the strongest predictor of derivative recognition knowledge when controlling for lemma frequency, proficiency and the interaction cognateness*proficiency. This direct effect of cognateness may have two interpretations: 1) that cognate words are also cognate regarding their morphological suffixes, and thus known better; 2) that since cognate words are better known in the form–meaning link, L2 learners progress earlier towards mastery of the derivative knowledge of this type of words. The latter interpretation would explain why this effect was found for the recognition mastery of derivatives but not its recall knowledge, since the former appears earlier in the developmental scale according to the results in Chapter 4.

As it was found for the previous components, the significant contribution of a words' frequency indicates that, as in the previous components, the higher the frequency

of a word the better the knowledge of recognition derivatives. Also, higher proficiency learners were significantly better in the recognition knowledge of derivatives.

The interaction between cognateness and proficiency is plotted in Figure 10.

Figure 10 Interaction between cognateness and proficiency in derivatives recognition.



This interaction effect suggests that the effect of proficiency is similar for cognate and non-cognate words in recognition of derivatives. Still, it seem that while the derivative recognition knowledge higher-proficiency learners have of cognates and non-cognates is similar, lower-proficiency learners seem to perform slightly better on cognates than non-cognates. This effect is not very large, though.

5.4.2.5. Multiple-Meanings Recall

Table 33 shows the comparative fit indices for the different fitted models.

Table 33 Comparison of model fit for the multiple-meanings recall aspect.

	Models	AIC	BIC
1	Full model	7,302.2	7,355.9
1.1	minus gender	7,301.0	7,348.7
1.2	minus word length	7,300.8	7,342.6
1.3	minus age	7,300.7	7,336.5
1.3	Cognates* Lemma Frequency	7,301.9	7,343.7
1.3	Cognates*Proficiency	7,271.1	7,312.9

Table 34 illustrates the best-fitting model for recall knowledge of multiple meanings. Age, word length and gender came out as non-significant predictors of this type of knowledge, and thus are not comprised in the model, without detriment of the model's overall goodness-of-fit. Similarly, no evidence of the cognateness*frequency interaction was found (z = .91, p = .36), and adding the interaction effect did not significantly improve the model fit (χ^2 (1) = .81, p = .37). Conversely, a significant interaction effect of cognateness and proficiency was found. ANOVA showed that adding this interaction did indeed improve the model significantly (χ^2 (1) = 31.63, p < .001). Thus, the final model included cognateness, lemma frequency and proficiency as predictors of recall multiple-meanings knowledge, as well as a significant cognateness*proficiency interaction effect.

The main significant effect of cognateness means that non-cognate words were known significantly worse than cognate words in terms of recall multiple-meanings knowledge. Or, in other words, cognates are known better in multiple-meanings recall knowledge. Again, cognateness proved to be the strongest predictor for this type of knowledge, followed by word frequency and proficiency. Thus, the higher the frequency of the words, the better the recall knowledge of multiple meanings, and the higher the learners' proficiency level, the higher their knowledge of this component. It is worth noting that the significance level of these effects on multiple-meanings recall are larger than in previous components (all at the p < .001 level). This suggests that the effect of all these predictors seems to be even more reliable for the multiple-meanings recall component.

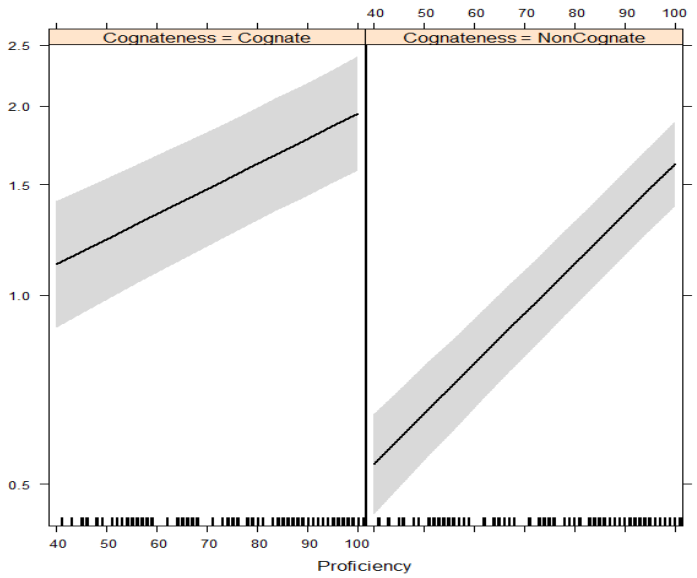
Table 34 Model of multiple-meanings recall knowledge.

	Multiple-Meanings Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-1.756	0.337	-5.201	***
Cognateness status	-1.100	0.192	-5.721	***
Lemma Frequency	0.153	0.030	4.947	***
Proficiency	0.009	0.001	7.535	***
Cognateness*Proficiency	0.009	0.001	5.635	***
Random Effects	***Variance***	***SD***		
Subject	0.00	0.00		
Item	0.07	0.27		

*** p < .001.

The interaction between cognateness and proficiency is plotted in Figure 11. This interaction shows a difference between the behaviour of learners of diverse proficiencies on cognates and non-cognates. That is: a larger effect of proficiency on knowledge of non-cognates than cognates was found at the multiple-meanings recall level. It seems that lower-proficiency learners recalled multiple meanings of cognate words much better than that of non-cognate words. It also indicates that higher-proficiency students knew the multiple meanings of cognate words slightly better than that of non-cognate words. Thus, there seems to be a facilitation effect of cognateness for both high and low-proficiency learners on multiple-meanings recall knowledge.

Figure 11 Interaction between cognateness and proficiency in multiple-meanings recall.



5.4.2.6. Multiple-Meanings Recognition

Table 35 shows the comparative fit indices for the different fitted models.

Table 35 Comparison of model fit for the multiple-meanings recognition aspect.

	Models	AIC	BIC
1	Full model	8,405.2	8,458.9
1.1	minus gender	8,403.2	8,450.9
1.2	minus age	8,401.8	8,443.6
1.3	minus word length	8,400.8	8,436.6
1.3	Cognates* Lemma Frequency	8,401.1	8,442.9
1.3	Cognates*Proficiency	8,388.0	8,429.7

Table 36 shows the final, best-fitting model for the multiple-meanings recognition component. Once again, age, word length, and gender did not show to be significant predictors of this type of knowledge, and are not included in the model. No evidence of a cognateness*frequency interaction was demonstrated ($z = 1.32$, $p = .19$), with no significant model fit improvement (χ^2 (1) = 1.68, $p = .20$).

As in the multiple-meanings recall aspect, cognateness, lemma frequency and proficiency proved to be significant predictors of recognition knowledge of multiple meanings. Cognateness showed to be the strongest predictor of this type of knowledge, suggesting that it is the factor that most influences better multiple-meanings recognition knowledge of cognate words. It was followed by frequency, with lower-frequency words being known worse than higher-frequency ones, and proficiency, where more proficient learners performed better in this component. The interaction between cognateness and proficiency came out as highly significant, and its inclusion proved to significantly improve the model (χ^2 (1) = 14.83, $p = .00$). Thus, the best-fitting model included cognateness, frequency and proficiency as fixed effects, and cognateness*proficiency as a significant interaction effect.

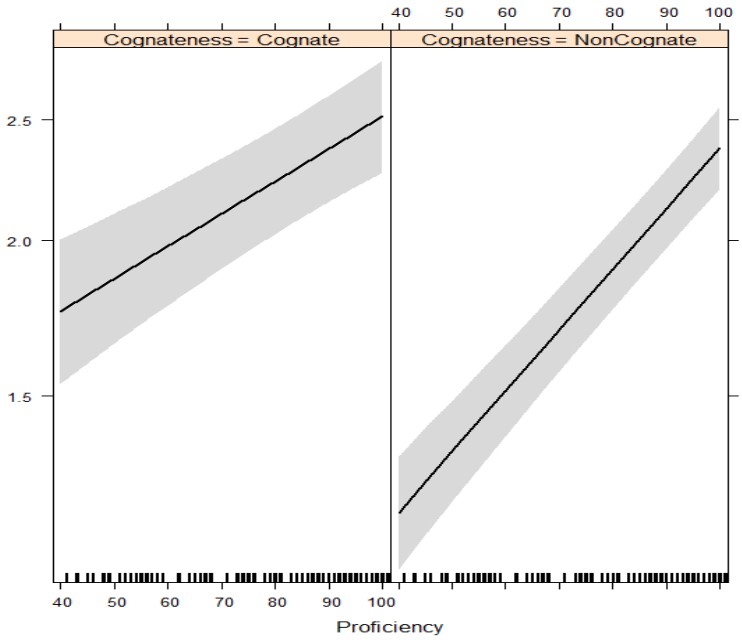
Table 36 Model of multiple-meanings recognition knowledge.

	Multiple-Meanings Recognition			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-0.588	0.177	-3.321	***
Cognateness status	-0.580	0.130	-4.463	***
Lemma Frequency	0.092	0.014	6.283	***
Proficiency	0.006	0.001	5.665	***
Cognateness*Proficiency	0.005	0.001	3.857	***
Random Effects	***Variance***	***SD***		
Subject	0.00	0.00		
Item	0.00	0.00		

*** $p < .001$.

The interaction between cognateness and proficiency is plotted in Figure 12. This interaction, as in the multiple-meanings recall aspect, suggests that lower-proficiency students recognised multiple meanings of non-cognates much worse than that of cognates. Higher-proficiency learners, on the contrary, knew cognates and non-cognate in the multiple-meanings recognition aspect to a similar degree.

Figure 12 Interaction between cognateness and proficiency in multiple-meanings recognition.



5.4.2.7. Collocation Recall

Table 37 shows the comparative fit indices for the different fitted models.

Table 37 Comparison of model fit for the collocation recall aspect.

	Models	AIC	BIC
1	Full model	2,612.2	2,665.9
1.1	minus gender	2,610.2	2,658.0
1.2	minus word length	2,608.3	2,650.0
1.3	minus age	2,606.5	2,642.3
1.3	Cognates* Lemma Frequency	2,607.4	2,649.2
1.3	Cognates*Proficiency	2,606.4	2,648.2

The best-fitting model for the recall collocations component is presented in Table 38. As in other aspects, the final model showed no significant effect of age, gender or word length on recall knowledge of collocations, and thus are not part of the model. No evidence of an interaction effect between either cognateness*frequency (z = -1.09, p = .28, no model fit improvement χ^2 (1) = 1.13, p = .29), or cognateness*proficiency (z = 1.48, p = .14, no model improvement χ^2 (1) = 2.13, p = .14) was found for this component. Thus, the best model fit included only cognateness, lemma frequency and proficiency as direct effect.

Table 38 Model of collocation recall knowledge.

	Collocation Recall			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-3.004	1.930	-1.556	.12
Cognateness status	-0.182	0.799	-0.228	.82
Lemma Frequency	-0.070	0.183	-0.383	.70
Proficiency	0.056	0.003	16.050	***
Random Effects	***Variance***	***SD***		
Subject	0.20	0.45		
Item	2.85	1.69		

*** p < .001.

Contrary to previous components, cognateness and frequency did not come out as significant predictors of recall collocational knowledge. The reason for this lack of an effect of cognateness and frequency may be due to the specific nature of these two factors in relation to collocational knowledge. The cognateness refers only to the cognate status of the target word, not of the two constituents in the collocation. Therefore, the results seem to indicate that cognate status of the node word is not sufficient to facilitate recall knowledge of the whole collocation. Cognateness of both elements in the collocation (i.e., total congruency) would have potentially yielded a significant effect, as evidenced by previous research (e.g., Peters, 2016). Frequency in this study refers to the lemma frequency of just the target word, therefore not including frequency information of the collocate. As in the cognateness case, frequency of only the node of the collocation is not enough to show a contribution in the knowledge of collocations recall. Had the whole collocation frequency been measured, an effect might have been expected, as some previous research has shown (González Fernández & Schmitt, 2015). Only proficiency showed a direct significant effect on recall collocational knowledge, suggesting that the learners' proficiency level seems to be the only factor that contributes to knowledge of recall collocations when controlling for cognateness and frequency. As expected, the higher the learners' vocabulary proficiency, the better their knowledge of recall collocations.

5.4.2.8. Collocation Recognition

Table 39 shows the comparative fit indices for the different fitted models.

Table 39 Comparison of model fit for the collocation recognition aspect.

	Models	AIC	BIC
1	Full model	2,348.4	2,402.1
1.1	minus word length	2,346.4	2,394.1
1.2	minus gender	2,344.4	2,386.2
1.3	minus age	2,343.1	2,378.9
1.3	Cognates* Lemma Frequency	2,344.5	2,386.3
1.3	Cognates*Proficiency	2,342.4	2,384.2

The final best-fitting model for recognition collocational knowledge is illustrated on Table 40. The results mirror those of its recall counterpart. Not surprisingly, age, word length and gender did not come out as significant, and are not included in this model. The model showed no interaction effect for either cognateness*frequency (z = -1.32, p = .19, no model improvement χ^2 (1) = 1.67, p = .20), or cognateness*proficiency (z = -.76, p = .45, no model improvement χ^2 (1) = .57, p = .45). Cognateness and frequency were not found to have an effect on this type of knowledge, probably due to the reasons described for the collocation recall model. Thus, the best-fitting model comprised fixed effects of cognateness, frequency and proficiency, but only the latter showed to significantly contribute to recognition knowledge of collocations, with higher-proficiency learners performing better in this aspect.

Table 40 Model of collocation recognition knowledge.

	Collocation Recognition			
Fixed effects	***β***	***SE***	***z***	***p***
Intercept	-1.631	1.081	-1.509	.13
Cognateness status	-0.590	0.438	-1.347	.18
Lemma Frequency	-0.031	0.100	-0.304	.76
Proficiency	0.052	0.004	13.334	***
Random Effects	***Variance***	***SD***		
Subject	0.32	0.57		
Item	0.80	0.90		

*** p < .001, ** p < .01, * p < .05

5.4.2.9. Summary of the results

Table 41 summarises the results of the models as far as the effects of proficiency, cognateness, frequency and word length are concerned.

Table 41 Summary of the mixed-effects modelling results.

Effects	FM Recall	FM Recog	Deriv Recall	Deriv Recog	MM Recall	MM Recog	Collo Recall	Collo Recog
Cognateness status	+	+	−	+	+	+	−	−
Frequency	+	+	+	+	+	+	−	−
Proficiency	+	+	+	+	+	+	+	+
Word Length	−	+	−	−	−	−	−	−
Interaction C*P	+	+	−	+	+	+	−	−

Note: plus sign (+) indicates a significant effect, minus sign (−) shows no significant effect.
FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.

Table 41 shows that for most of the word knowledge components the results pointed in the same direction: word length is not a good predictor of overall vocabulary, cognateness and frequency are general indicators of vocabulary knowledge, with cognate words and frequent words typically being known better, and proficiency is the most widespread predictor of word knowledge.

Proficiency is the common factor that predicts vocabulary knowledge across all aspects, indicating that higher proficiency leads to better knowledge in all aspects. Knowledge of all the components, except collocational knowledge, seems to be strongly influenced by the frequency of the words' lemma. This suggests that frequency contributes significantly to how well words are known in the different components, with higher-frequency words generally being known better. Knowledge of collocations seems to be more influenced by other factors than frequency when only the target word' frequency is accounted for. Cognateness status affects knowledge in most of the word knowledge components, emerging as the strongest predictor of knowledge in these aspects. That is, cognate words are generally better known that non-cognates across different word knowledge aspects. A lack of this cognateness status was found in recall derivative knowledge and collocational knowledge overall. The finding that cognateness does not have an effect on collocational knowledge is not surprising, since the cognate words are only the nodes of the collocation, and they do not necessarily imply a cognate collocate. The lack of cognate effect on recall derivative knowledge suggests that knowing the cognate target word does not imply knowing its derivatives.

5.5. Discussion

The study aimed to investigate whether various inter and intralexical factors predict the odds that Spanish EFL learners know different word knowledge components in recognition and recall. In particular, the effect of the following factors on word knowledge will be discussed: cognateness, frequency, proficiency and word length. Answering the first and second research questions, all these factors seem to contribute to vocabulary knowledge as a whole, although their influence varies across the individual word knowledge components.

5.5.1. Effect of target word cognateness

Research has shown that cognates are easier to acquire than non-cognates for L2 learners, and thus cognateness status has been claimed to be one of the best predictors of learning (Elgort, 2013; Laufer & Nation, 2011; Lubliner & Hiebert, 2011; Willis & Ohashi, 2012). The fourth research question in this study asked whether cognateness status has an effect on all the word knowledge components. The nature of cognate words implies that their form and meaning are similar in two languages, but does not suggest that they are similar in other aspects (e.g., derivatives or multiple-meanings). Therefore, it could have been expected that cognateness was a contributor of knowledge of the form–meaning link but not the other components. Descriptive statistics showed that cognate words were known better, on average, than the non-cognate words in most of the components. The results of the mixed-effects models confirm this trend and show an overall effect of cognateness status on vocabulary knowledge. This suggests that learners are able to transfer vocabulary knowledge form their mother tongue to recognising and recalling overall vocabulary in the L2. However, this effect varied across components.

Previous research has shown a facilitation effect of cognateness status on the form–meaning link aspect, both in recognition and recall, with learners providing more accurate responses to cognates than non-cognates in this aspect of knowledge. (e.g., Cunningham & Graham, 2000; Elgort, 2012; Laufer & Ravenhorst-Kalovski, 2010; Méndez Pérez et al., 2010; Rosselli et al., 2012). My findings are in line with this research, and show that, indeed, cognateness status had a highly significant direct effect

on knowledge of the recognition and recall form–meaning link, after controlling for variables such as proficiency and frequency. In fact, cognateness proved to be the strongest predictor of this type of knowledge. This indicates that whenever L2 learners of cognate languages are being tested on their knowledge of the form–meaning link of words, cognateness status would be the best indication of the likelihood of knowing a word over the frequency of that word. This cognateness facilitation has two main consequences on the testing of vocabulary size (generally operationalised by knowledge of the form–meaning link). On the one hand, when cognate words are included in these tests, it may lead to an overestimation of the vocabulary size of those L2 learners if the proportion of cognates tested is higher than that between the two languages (Elgort, 2013). Thus, the results would need to be interpreted in light of the percentage of cognate words included. On the other hand, it has been suggested that the frequency principle to estimate knowledge might be problematic in certain situations, as vocabulary development is not only reflected on the knowledge of low-frequency words, but also on the extended use of frequent words or awareness of other types of knowledge (Fitzpatrick & Clenton, 2017). In this sense, correct identification and recall of cognate words would also demonstrate better vocabulary knowledge, and should be part of vocabulary size tests.

Little research has explored how cognateness status affects knowledge of derivatives and multiple meanings, and thus there is no evidence to believe that it would have an effect on these types of knowledge. However, my analysis shows that the cognateness of a target word predicts knowledge of derivatives (in the recognition mastery) and multiple meanings (both recognition and recall). This interesting finding demonstrates that cognateness of target words does not only influence knowledge of the form–meaning link, but also of deeper word knowledge types such as derivatives and multiple meanings. This may indicate that when a word shares a similar form and the same first meaning in two languages it is likely to share also certain morphological characteristics and multiple meanings, which would make it easier for L2 learners of that cognate language to acquire more senses or derivatives of such words. For example, the different meanings of the cognate word *employ* are similar both in English and Spanish, which explains the effect on recall and recognition knowledge of multiple meanings. However, this is not always the case for all cognate words. Some words have very different second and third meaning in Spanish and English (e.g., *bank* as ‘monetary

institution' → *banco*; '*bank* of the river' → *orilla*; 'a plane to *bank'* → *ladear*; or '*redeem* past behaviour' → *redimir;* '*redeem* a voucher' → *canjear*; '*redeem* a debt' → *liquidar*). Moreover, even in the cases where both languages share the various senses, these may have different frequencies, or the specific meaning could also be represented by other words, sometimes even more frequent than the cognate form. For example, in Spanish the meaning of *employ* as 'providing a job for somebody' is more commonly conveyed with the word *contratar,* instead of *emplear*, which is the cognate word for that meaning. This interpretation is admittedly just a speculation, and further research will need to determine the reasons for cognate effect on knowledge of multiple meanings.

As for knowledge of derivatives, some research has found a relationship between L1 derivational knowledge and L2 cognate vocabulary knowledge, so that the former facilitates recognition of the latter (Hancin-Bhatt & Nagy, 1994; Nagy et al., 2006). For example, Ramírez et al. (2013) studied how recall and recognition knowledge of derivatives by 90 Spanish speakers affected their L2 English vocabulary of cognate and noncognate words. They found that derivational knowledge was directly related to English cognate vocabulary, but not to English non-cognate vocabulary, and concluded that knowledge of cognates facilitates the transfer of Spanish derivational knowledge to acquisition of English vocabulary. Dressler et al., (2011) found that L1 Spanish-L2 English children recognised the cognate roots of derived words in English, showing the connection between cognateness and recognition derivational knowledge. My findings provide evidence for the influence of cognates on the recognition knowledge of derivatives, and suggests a possible effect of L1 knowledge on the better recognition of derivatives. For example, the word *character* is similar to the Spanish in the noun (*character* vs. *carácter*), verb (*characterise* vs. *caracterizar*), adjective (*characteristic* vs. *característica*) and follows the morphological rule to apply the most common suffix to create an adverb (*characteristically* vs. *característicamente*). This case supports the facilitative cognate effect on recognition of derivatives. That being said, this almost-parallel derivative correspondence does not hold for all the cognate words (e.g., *freshness* vs. *fresco, freshen* vs. *refrescar*, *fresh* vs. *fresco, freshly* vs. *frescamente*).

The above counter-examples suggest that cognate words cannot be considered perfect 'cognates' in terms of derivatives and multiple meanings (i.e., always sharing

the same or most equivalent derivational suffixes and/or meanings). Considering this, a reason for the predictive effect of cognateness on recognition derivatives and multiple-meanings knowledge may be that cognate words, due to their nature, are learnt easier, quicker and better in the form–meaning link, allowing learners to devote their cognitive resources to noticing and focusing on other aspects of the words. As a result, they would progress quicker on the implicational scale towards more advanced types of knowledge, such as that words' multiple meanings or derivatives.

Conversely to the results on the form–meaning, recognition derivatives, and multiple-meanings aspects, the cognateness status of the target word did not have an influence on the derivatives recall or on either the recognition or recall of collocational knowledge. The lack of cognate effect on the recall knowledge of derivatives suggests that, when learners have to decide on the appropriate suffix for a word, they have difficulties regardless of whether it is a cognate or not. This confirms the intuition that even if learners know suffixes, they struggle in selecting the appropriate one for a particular word, and supports previous research demonstrating the complexity of derivational knowledge, particularly in the recall form (Kieffer & Lesaux, 2012a; Nagy et al., 2006). Regarding collocations, the lack of an effect of cognateness contradicts previous research showing an influence of the L1 on collocational acquisition and finding generally better knowledge of cognate (congruent) collocations (e.g., Gyllstad & Wolter, 2015; Peters, 2016; Wolter & Gyllstad, 2011). The reason for these opposing results rely on the item being considered as a cognate. When the whole collocation as a unit is understood as a cognate or congruent, these items are found to be better known than those which are non-cognates/incongruent. In contrast, my results show that when a target word/node of a collocation is a cognate in the learners' L1 it does not mean that the whole collocation is going to be a cognate/congruent. Therefore, the cognateness status of the target word is not sufficient to influence collocational knowledge.

Finally, the analysis shows that for all those components where cognateness came out as a predictor of knowledge (i.e., overall form–meaning link, overall multiple-meanings, and derivatives recognition) there was also evidence of a significant interaction between proficiency and cognateness. Previous research by Duñabeitia, Cholin, Corral, Perea, & Carreiras, (2010) had proposed that although cognateness automatically activates a lexical connection in bilinguals, this activation is modulated by the proficiency level of the participants. My study shows that this, indeed, the

likelihood of correct responses to cognate words was conditioned by the learners' proficiency in five of the eight vocabulary components tested: recall and recognition of the form–meaning link, recognition of derivatives, and recall and recognition of multiple meanings. For all of these components, the interaction meant that lower-proficiency learners answered cognate words better than non-cognate words in the specific type of knowledge, while higher-proficiency learners knew them to a similar degree (except for form–meaning recall where they seemed to perform better on non-cognates).

Some research has claimed that lexical knowledge of cognates is not only related to the proficiency of learners, but also to the frequency of the words (Dressler et al., 2011). Chen et al. (2012) also acknowledged the frequency of the cognate word as a factor that influences the effect of cognates on vocabulary knowledge, suggesting a potential interaction effect between cognateness and frequency. They claim that, since most low-frequency English words are high-frequency Spanish words, it is possible that knowledge of low-frequency English words by L1 Spanish learners is influenced by cognateness status. However, my study shows no evidence of such an interaction, suggesting that cognates are generally better known regardless of their frequency in English.

The fifth research question considered whether frequency was a stronger predictor of vocabulary knowledge than cognateness, or vice versa. The analysis of a predictor effect of cognateness status on the knowledge of the eight word knowledge aspects indicated that, when cognateness showed a significant effect, it was always the strongest predictor of vocabulary knowledge after controlling for the lemma frequency of the target items and the proficiency of the participants. Previous research has also found cognateness to be the best predictor of vocabulary knowledge over frequency. For example, Vidal (2011) studied how predictability from word forms (i.e., cognateness), proficiency, corpus frequency and frequency of occurrence affected the incidental acquisition of vocabulary. She found that cognateness was the best predictor of lexical acquisition, accounting for up to 57% of the variance. These findings suggest that for L2 learners of cognate words, although frequency is still a significant indicator of vocabulary knowledge, it is the cognateness status of words which best predicts their knowledge.

Overall, this study confirms the notion that L2 words that remind us of corresponding L1 words due to form and meaning similarities seem to be generally better acquired and retained than those that are different in both languages (De Groot & Keijzer, 2000; Vidal, 2011). It supports the previously demonstrated effect of cognateness on knowledge of the form–meaning link, and provides initial evidence for this facilitative effect of cognateness on other types of word knowledge, particularly multiple-meanings and derivative knowledge. Also, my results seem to provide further support for theories of crosslinguistic transfer (e.g., the linguistic interdependence hypothesis by Cummins (1979)), which hypothesise that the L1 metalinguistic and literary skills of bilinguals can be transferred to their L2 learning (Koda, 2008; Lubliner & Hiebert, 2011). Therefore, when examining and interpreting vocabulary acquisition in L2s, we need to better understand the interface between the learners' experience with the L2 and their knowledge of the L1 (Wolter & Yamashita, 2017).

5.5.2. Effect of target word frequency

The frequency of words is widely accepted as one of the best predictors of lexical knowledge, with higher-frequency words commonly being known better. However, this claim is chiefly based on research looking at the form–meaning link. Thus little evidence exists about how frequency influences knowledge of other components (i.e., multiple-meanings), or the findings on its influence are inconclusive (e.g., its effect on collocations). The results of this study show that the lemma frequency was clearly a strong predictor of knowledge across all the word knowledge components (p <.01-.001), except collocational knowledge. The positive frequency coefficients confirm that words of higher-frequency are more likely to be known both in the recognition and recall aspects of form–meaning, derivatives, and multiple meanings.

The finding that corpus frequency positively affects knowledge of the form–meaning link is not surprising, since much previous research has evidenced this pattern (e.g., Diependaele et al., 2013; Vidal, 2003, 2011). It demonstrates that frequency is a consistent indicator of form–meaning knowledge, and thus we can predict the likelihood of learners' knowing certain words based on this count. My findings confirm the value of employing frequency to inform the development of both vocabulary size tests (Vocabulary Size Test (Nation & Beglar, 2007), Vocabulary Levels Test (Schmitt et al.,

2001), New Vocabulary Levels Test (McLean & Kramer, 2015)), and word lists (e.g., *General Service List* (West, 1953), *Academic Word List* (Coxhead, 2000), *New General Service List* (Brezina & Gablasova, 2015; Browne, 2013), *New Academic Vocabulary List* (Gardner & Davies, 2013)) from the earliest days.

Although there are some claims that frequency affects also morphological knowledge (Crossley et al., 2010), limited research has investigated this effect on derivative knowledge. One study by Saigh (2015) examined the relationship between frequency and knowledge that L2 learners of English had of derivatives recognition and recall. She found that words' frequency, understood as frequency of all accepted derivative forms of a target word, significantly correlated with their knowledge on both tests, explaining 41% and 31% of the variance respectively. My study provides evidence of the effect of the target words' lemma frequency on both recall and recognition derivative knowledge, indicating that learners know better the derivatives of frequent words than those of infrequent words. A reason for this finding is that, since frequent words are learnt in the form–meaning link earlier, learners can progress towards deeper knowledge of that word (including derivatives) sooner.

Multiple meanings is another aspect for which the effect of frequency is still uncertain. A study by Crossley et al., (2010) found a significant positive correlation between frequency and multiple meanings scores of L2 learners ($r = .59$, $p < .001$), confirming their hypothesis that frequency and multiple meanings were related. Their findings suggest that when learners use more frequent words, they also use more senses of those words. In other words, the more frequent words are, the more senses L2 learners are likely to know of those words. My study provides support for this claim, suggesting that the frequency of target words may be able to inform about the likelihood of knowing multiple meanings of that word.

Finally, no effect of frequency was found for collocational knowledge. This fact highlights once again the different behaviour of collocations as opposed to other word knowledge aspects, and complements previous research that found relationships between whole collocation frequency and knowledge of L2 collocations (González Fernández & Schmitt, 2015; Vilkaitė & Schmitt, 2017; Wolter & Gyllstad, 2013). The reason for this finding might be, however, that I accounted for only the frequency of one of the constituents in the collocation, and this might not be sufficient to find evidence of the frequency effect on collocational knowledge. My result correlates with

previous findings by Zhang and Yang (2016), who found no effect of the target word's frequency on receptive knowledge of associations and collocate. This lack of frequency effect was also evidence by Macis and Schmitt (2017a, 2017b) when the whole collocational frequency was accounted for. Yet, psycholinguistic research has found frequency effects on collocational processing (e.g., Conklin & Schmitt, 2008; Wolter & Gyllstad, 2011). A recent study has suggested that L2 learners can and do attend to both word and collocational frequency, but that, as learners become more proficient they seem to shift from relying on word-level frequency to relying on collocational frequency (Wolter & Yamashita, 2017). This greater reliance on word-level frequency by less proficient students is supported by Nguyen and Webb's (2017) finding that the node word frequency was the strongest predictor of receptive collocational knowledge over collocational knowledge for their relatively low proficiency learners of English. Considering this, it could be that the generally good proficiency of my participants led to the lack of effect of word frequency on collocational knowledge, since they would have already moved onto reliance of collocational frequency. Until more research is conducted in this topic, though, my study seems to add to the list of collocational studies that find inconclusive results regarding the effect of frequency on this component (e.g., Pellicer-Sánchez, 2017; Szudarski & Carter, 2016).

Overall, the study provides support for the relationship between frequency and general vocabulary knowledge. It also demonstrates that the importance of frequency as a predictor of familiarity with individual words extends to other word knowledge components, not just the form–meaning link. Moreover, it confirms previous claims that even though frequency is one of the best contributors to vocabulary knowledge, it is not the strongest predictor when controlling for cognateness status (Laufer, personal communication; Vidal, 2011). Thus, whilst both frequency and cognateness affected knowledge of overall vocabulary, the effect of cognateness was generally stronger, suggesting that cognateness may be a better predictor of vocabulary acquisition than frequency.

5.5.3. Effect of proficiency

Proficiency in this study was conceptualised as vocabulary level, and measured by means of a VLT composite score. Participants showed to have a reasonably good

vocabulary proficiency, as shown by 68% right answers on average in the VLT. More importantly, although they did not master even the 2K individual section (average score 83%), participants knew almost half of the items in the 10K section (43%), confirming that cognates played a role in their vocabulary knowledge (Malabonga et al., 2008). In fact, in this study, the participants' proficiency was found to be the only common predictor of knowledge across all the vocabulary components. This suggests that, although proficiency is not the best predictor after controlling for frequency and cognateness, it would be the most widespread indicator of overall vocabulary knowledge components.

My findings are consistent with previous research showing the robust effect of proficiency on L2 knowledge of the form–meaning link (Lee & Pulido, 2017; Montero Perez et al., 2014; Vidal, 2011; Zahar et al., 2001). Proficiency was also found to be a predictor of collocational knowledge. In fact, it was the only significant predictor of this type of knowledge. This suggests that the vocabulary size of a person seems to be a very clear indicator of their knowledge of collocations, as previous research has found (Peters, 2016; Vilkaitė & Schmitt, 2017). The effect of proficiency on other components such as knowledge of multiple meanings or derivatives has been the subject of fewer studies. As for the former, some researchers had inferred a relationship between proficiency and knowledge of a words' multiple meanings (e.g., Crossley et al., 2010). My study provides direct evidence that proficiency is indeed a predictor of multiple-meanings knowledge, so that the larger the vocabulary size, the greater the likelihood of knowing multiple senses of a word. In the case of derivatives, Saigh (2015) found very strong correlations between vocabulary size and recognition and recall knowledge of derivatives (r =.94 and .93 respectively). Similarly, Hayashi and Murphy (2011) found a strong correlation between derivatives, form recall knowledge and the vocabulary size (r =.84). Although with more moderate correlations, Schmitt and Meara's (1997) study confirms this trend, showing correlations of between r =.27-.41 between vocabulary size and recognition and recall derivative knowledge. The positive direct effect of proficiency on derivative knowledge supports their relationship and provides evidence that proficiency can be used as an indicator of derivative mastery.

In addition, proficiency was part of a significant two-way interaction with cognateness for most components (form–meaning link, multiple-meanings and recognition derivatives). This relationship between cognate knowledge and proficiency

had been previously evidenced for Spanish EFL learners (Malabonga et al., 2008). Analysed further, these interactions suggested that the facilitative effect for cognate words depended on learners' proficiency. It seems that for participants with higher proficiency, there was almost no effect for cognates, as the participants knew both cognate and non-cognate words regardless. But for participants who had smaller vocabularies, cognates were much better known that non-cognates; that is, the cognate facilitation increased in lower proficiencies. This finding that proficiency moderates the cognate effect in most components could indirectly suggest that an increase in vocabulary size may lead to an increase in knowledge of non-cognates. That is, more advanced learners are filling in the non-cognates that lower proficiency learners do not know. This is connected with previous research finding that longer time spent in the L2 environment led to better non-cognate knowledge by L1 Spanish learners (Chen et al., 2012), suggesting that proficiency and amount of exposure to the L1 explain cognate knowledge (Malabonga et al., 2008).

To summarise, my study shows an overall effect of proficiency (i.e., vocabulary size) on general word knowledge components. The coefficients of the VLT scores suggest that the larger the vocabulary proficiency of learners, the better their knowledge in all the word knowledge components. This finding is not surprising, as the increase in one's vocabulary size is related to an increase in depth of vocabulary knowledge (Schmitt, 2014).

5.5.4. Effect of word length

It has been suggested that longer words are more difficult to learn because there is more to remember (Nation & Webb, 2011). Following this claim, a negative effect between word length and word knowledge would be expected. My results showed a lack of a significant effect of this factor on knowledge of most vocabulary components, suggesting that word length is not a clear predictor of overall knowledge. Moreover, the fact that no effect was found for most components could be an indirect indication that word length is difficult to isolate from other factors such as morphology, or word frequency (Laufer, 1997; Milton, 2009). This way, the effect of word length may have been comprised in other factors such as word frequency.

The only word knowledge component for which a significant effect of word length was found was recognition knowledge of the form–meaning link. This is not surprising, considering that word length has been found to be a significant factor in word processing and reading (e.g., Kliegl, Grabner, Rolfs, & Engbert, 2004; Vilkaitė, 2016) and form–meaning recognition is a highly related component to reading behaviour. However, the findings suggest that the longer the target word, the better they were known in the recognition form–meaning link aspect. This result seems surprising, as it contradicts not only the intuition but also previous research showing that shorter words are generally recognised (Barclay, 2017; Campoy, 2008) and produced (Barclay, 2017) better than longer words at the form–meaning link level. Nevertheless, this finding is in line with recent research by Peters (2016) who found that longer collocations were better known at the form–meaning recognition level. A potential reason for this result is that, while longer words are more demanding to acquire for learners, once acquired they may be remembered better at this recognition level. Although Barclay (2017) found a lack of word length effect on word retention, further research is needed to explain these findings.

5.5.5. Pedagogical implications

Obtaining a better understanding on the factors that affect vocabulary knowledge can help teachers anticipate the potential learning burden of words and inform the steps needed to improve lexical instruction. Frequency has long been accepted as a predictor of vocabulary knowledge, and thus many teaching and testing materials are based on the frequency principle (e.g., word lists, graded readers, vocabulary size tests). My study provided further evidence that these frequency-based materials should be applied to the learning and teaching of not only the form–meaning link, but also multiple meanings and derivatives. Also, I found that frequent words are more likely to be known across different components of size and depth of vocabulary. Thus, if deep mastery of words from a wide range of frequencies is the target, practitioners should provide learners with incidental opportunities to develop various types of knowledge of frequent words, while devoting the instruction time on developing deeper knowledge of less frequent words. Moreover, pedagogy should also focus on developing awareness of frequency constraints on use of words, not only the

differentiation between frequent and infrequent words. This knowledge would enable learners to choose the most appropriate word for each situation, instead of using the frequent words they are familiar with in all contexts (Wilks, 2014).

The fact that cognateness has been shown to be the strongest predictor of knowledge for most word knowledge components suggests that teachers should take advantage of the learners' L1 for teaching the L2. There is evidence that Spanish learners of L2 English can use their L1 vocabulary knowledge to identify, understand and use L2 vocabulary (e.g. August et al., 2005; Hancin-Bhatt & Nagy, 1994; Jiménez, García, & Pearson, 1996). For this reason, researchers advocate that previous language knowledge, literacy skills and metalinguistic knowledge should be used in the classroom, specially drawing on the knowledge of cognate or equivalent expressions as a means of learning new vocabulary (Carlo et al., 2004; Molnár, 2010; Rodríguez, 2001).

Yet, research suggests that L2 learners do not always employ the ability to use cognates as vocabulary learning strategy, probably due to their lack of awareness. Thus, given that the lack of cognate awareness hinders the opportunities for transfer of L1 knowledge to L2 knowledge, cognate awareness and cognate-based instruction must be developed as a vocabulary learning technique. Explicit cognate instruction and training in identifying and recognising cognates has been frequently cited and proven successful as a strategy to enhance L2 word knowledge (Bravo et al., 2005; Carlo et al., 2004; Dressler et al., 2011; García, 1991; Lubliner & Hiebert, 2011; Otwinowska-Kasztelanic, 2009; Proctor & Mo, 2009). Once learners develop a cognate recognition ability, they need to be trained in analyzing their meaning based on the context in which they appear, in order to avoid false cognates' effect and infer meaning successfully. Dressler et al., (2011) showed that direct instruction in cognate recognition and identification lead to better knowledge of cognates and this resulted in better inferring from context abilities in Spanish-English bilingual children. Likewise, Carlo et al. (2004) found that the Spanish learners of English in a word-inference skills intervention group experience greater vocabulary growth than their peers who did not receive instruction. Considering the substantial number of cognates shared by English and Spanish, rising awareness of the existence of cognates and teaching cognates, maybe in the form of word lists, would be expected to greatly facilitate learning new words (size) and a lot about those words (depth) in English. Moreover, cognate knowledge has also been found to benefit from

incidental learning, both from reading and listening (Vidal, 2011). Thus, once awareness of cognates has been highlighted, and these words have been deliberately taught, teachers can employ reading and listening tasks to enhance their learning and improve their inferential skills (in the case of false cognates).

5.5.6. Limitations and directions for further research

There are inevitably some limitations to the study. First of all, the number of cognate and non-cognate words included in the analysis was not equivalent. The reason for this was the strict criteria required for the selection of the target words, which made it very difficult to match these words. Although mixed-effects modelling is a powerful technique that can account for this difference, having a more balanced proportion of cognates and non-cognates would have provided the study with more control. Another issue regarding the target words is that the two types of items (i.e., cognates vs. non-cognates) were not perfectly matched on morphological structure (word class) or frequency. Although I ensured that both types of items had a range of frequencies (cognates: 1K-6K; non-cognates: 1K-9K), it was difficult to match them perfectly on any dimension due to their inherent differences (cognate words in English tend to be low-frequency academic words, while non-cognate words are generally high-frequency, daily vocabulary). Controlling for these factors in a stricter manner would provide future research with a more powerful design.

Secondly, even though proficiency was conceptualised in this study as vocabulary size, the measurement instrument used to assess it (VLT) has not been developed as an actual measure of size, but to provide an estimation of the participants' vocabulary profile (Kremmel & Schmitt, 2017). Therefore, although the VLT was considered an appropriate measure in this study to provide a general vocabulary proficiency estimate, it is important to bear in mind that it should not be interpreted as the actual vocabulary size of participants. Moreover, only the L2 vocabulary proficiency was measured in this study. The participants' L1 proficiency is also likely to have an effect on their knowledge of cognates and of the different word knowledge components. Future research may want to include a more standardised measure of L2 vocabulary proficiency, while also controlling for the L1 proficiency of the participants.

Finally, the results show that cognateness status influenced the knowledge that the Spanish participants have of the different word knowledge components, suggesting that this is the case also for L2 learners of any cognate language. Therefore, one question that needs to be addressed is whether the order of the implicational scale presented in Chapter 4 has been affected by the cognate nature of the participants' L1. In order to explore this, research is needed that replicates the findings in the study presented in Chapter 4 and the present study with a population of non-cognate language. If results remain the same regarding the implicational scale, then we can claim that a universal order of word knowledge components might exist and could be generalised to all languages. Conversely, if the order varies, it could suggest that the cognate effect is so strong that cognate and non-cognate languages follow different orders of overall vocabulary acquisition. This gap will be addressed in Chapter 6.

5.6. Conclusion

The present study investigates whether various interlexical and intralexical factors affect different word knowledge components and to what extent. Mixed-effect modelling analysis was employed to explore the role of cognateness, proficiency, frequency and word length on overall vocabulary knowledge. The results displayed differences and similarities in the role played by these factors on each of the tested components: recognition and recall knowledge of form–meaning link, derivatives, multiple meanings and collocations. Cognateness came out as the strongest predictor of overall vocabulary knowledge, with cognate words being known significantly better than non-cognates. Frequency was found to be the second best indicator of L2 knowledge across all the word knowledge components, except collocations. Unsurprisingly, word length proved to have a significant effect on recognition of the form–meaning link, although, unexpectedly, longer words were found to be known better than shorter words. Finally, proficiency came out as the only common factor contributing to the knowledge of all vocabulary components, although its contribution was less prominent than that of cognateness and frequency. Overall, this study demonstrates that the factors that had previously been found to have an effect on the form–meaning link aspect of words also extend their influence to other deeper word knowledge components, providing support for the claim that the complexity of language

learning very likely lends itself to the influence of multiple factors (Ellis & Larsen-Freeman, 2006).

Chapter 6

The vocabulary knowledge construct and the relationships among its components: A Structural Equation Model

Describing an order of difficulty, and thus a pattern of acquisition, of the word knowledge components is an important step towards a better understanding of vocabulary knowledge. However, it is not sufficient to explain the overall structure of this complex construct. Although vocabulary knowledge is theoretically described, interpreted and accepted as a multidimensional construct, this multidimensionality has never been empirically investigated. Thus, it is still unclear how the different word knowledge components contribute to the overall vocabulary knowledge construct, and whether some predictive relationships between them can be specified. The results from Study 1 in Chapter 4 suggest that the vocabulary knowledge components interrelate so highly with each other that it may not be possible to consider them independent concepts. This raises the question of whether vocabulary is actually multidimensional in practice. Study 3 (in the present chapter) was designed to address this question. This chapter reviews the results from multicomponent studies involving vocabulary, introduces the statistical technique required to examine the structure of the vocabulary construct and reports on a study that explores the relationships among word knowledge components in order to provide empirical evidence of the nature of vocabulary knowledge.

6.1. Background of the study

6.1.1. Multidimensional interpretation of vocabulary knowledge

The vocabulary knowledge construct is generally accepted as comprising multiple word knowledge dimensions within it (Milton & Fitzpatrick, 2014; Nation,

2013; Read, 2000; Schmitt, 2014). As a result, the most widely followed approach to vocabulary knowledge is the components approach, which divides vocabulary into its component parts. Paul Nation (2013) is the best representative of this approach, due to his framework describing the various word knowledge aspects. This framework is the most comprehensive and the one referred to by most current vocabulary scholars in the description of vocabulary (see Table 1, Chapter 2). The components specified in this conceptualisation are considered to be interrelated with one another. However, most researchers have treated them as independent from each other, and very little research has attempted to empirically show the interrelationships between them, leaving this important issue unanswered.

As discussed in Chapter 2, the great complexity of the vocabulary knowledge construct with its many components makes it problematic to apply in research, which has led most researchers to measure the different aspects independently. Even in the few occasions in which multiple components have been examined concurrently, the results are based on a low number of participants, few target words, and, more importantly, interpretation of mean scores and correlations only (see Section 2.2.3 in Chapter 2 for a more detailed discussion of these studies and their limitations). The great intricacy of the vocabulary knowledge construct demands more rigorous methods of analysis that can inform about the relationships among word knowledge components, their contribution to overall vocabulary knowledge, and the multidimensionality of the construct. The next section reviews some studies that have adopted this more sophisticated approach, as well as their limitations for the understanding of the vocabulary knowledge construct.

6.1.2. The use of SEM in the examination of vocabulary knowledge

The multifaceted structure of vocabulary knowledge requires it to be explored by means of statistical techniques capable of measuring latent variables. Latent variables represent constructs that are comprised of multiple aspects. Thus, vocabulary, as a multidimensional concept, necessitates a method that can contain and analyse its different components within a latent variable. This method is called Structural Equation Modelling (SEM) (a comprehensive description of this technique is provided in Section 6.2 in this chapter).

SEM is a commonly employed method of analysis in the social and behavioural sciences, and thus it has also been applied to the study of language acquisition. In a review on the use of SEM in applied language testing and learning, In'nami and Koizumi (2011) found that SEM was most often used to investigate language testing, particularly test-taker variables (e.g., learner's strategy use (Purpura, 1997)) and test/task-related variables (e.g., trait/test structure (Schoonen et al., 2003)). However, as they assert, "all topics in the language testing and learning fields would merit further investigation using SEM" (In'nami & Koizumi, 2011, p. 270). This is particularly true for the vocabulary field, as the construct has not been largely explored by means of SEM and our understanding of it would certainly benefit from this examination.

There are a few studies that have investigated some aspects of vocabulary employing this refined statistical method. Their main focus could be divided into: the contribution of learning strategies to vocabulary, the contribution of vocabulary to reading comprehension, and the relationships between vocabulary and other lexical aspects.

6.1.2.1. Learning strategies and vocabulary knowledge

One of the first studies that investigated the influence of different learning motivational strategies to vocabulary knowledge by means of SEM was Tseng and Schmitt (2008). They analysed the strategic behaviour and vocabulary knowledge of 259 Chinese university learners by employing a vocabulary learning strategies questionnaire, a written form recall test to measure vocabulary size, and a receptive polysemy and collocation test to assess depth. They found that motivated vocabulary knowledge is a systematic and cyclical process, and that it requires going through several motivational stages before it is acquired. Building on the previous study and the understanding of L2 vocabulary learning as a cyclical process, Ma (2015) suggests a model in which successful vocabulary knowledge goes through four sequential stages: perceiving the word form, accessing the word meaning, mapping the word form with meaning, and using the word. This author also understands vocabulary knowledge as size (measured by the 3K and 5K VLT) and depth (measured by half of the WAT test). Assessing the same learner population, the results showed that all the motivational stages contributed directly and indirectly to vocabulary knowledge, although the

contribution of form was much lower than that of meaning or use. More importantly for the purpose of the present study, Ma (2015) found that the size measure contributed to vocabulary knowledge more than the depth measure (67% vs. 45%), although both showed to be strongly connected to the vocabulary construct.

6.1.2.2. Vocabulary knowledge and reading comprehension

SEM has also been used to examine the effect of vocabulary on reading comprehension, particularly comparing it with the contribution of grammar. Considering vocabulary knowledge as comprised of size (measured by the VLT) and depth (measured by the WAT), Zhang (2012) found that vocabulary knowledge was a stronger predictor of L2 reading comprehension than grammar for 190 advanced Chinese EFL learners. Kremmel et al. (2017) also studied the impact of syntactic, vocabulary and phraseological knowledge on the reading ability of 418 Austrian EFL learners. They employed only one measure for each of these construct. Vocabulary knowledge was assessed by means of the DIALANG Advanced Vocabulary Test (Alderson, 2005), which measures receptive and productive knowledge of form–meaning, collocations, derivations and associations (i.e., depth). Phraseological knowledge was measured by means of a multi-word expression, multiple-choice recognition test. They found that phraseological knowledge was the strongest contributor to reading comprehension (β = .57), followed by vocabulary knowledge (β = .29). However, they discovered a .90 correlation between the phraseological construct and the vocabulary construct, which in fact, unlike the authors' conclusion, indicates that the two factors cannot be empirically distinguished from each other, and thus they should be better considered one construct comprising different types of lexical knowledge.

This strong contribution of vocabulary knowledge to reading comprehension has also been studied on NS of English. Assessing the effect of various language skills on reading, Babayiğit (2015) found that the construct which comprised vocabulary knowledge within it (measured by means of the PPVT) was the best contributor to reading comprehension both in English as L1 and L2, although the effect was stronger for the L2 group. In a study only with NS, Tannenbaum et al. (2006) understood vocabulary knowledge as comprised of three components: breadth, depth and fluency,

and measured this knowledge on 203 third-grade children to assess their effect on reading ability. They employed two different standardised measures for each of the three dimensions. Breadth of word knowledge involved a picture multiple-choice form–meaning link test. For depth, they used a multiple-meanings recall test, plus an attributes test, in which participants had to name different attributes of a noun. For fluency, students had to use as many words as possible in sentences per minute, and to name as many items as possible in different categories, with 10 seconds for each category. They found that only knowledge of vocabulary breadth contributed to reading comprehension significantly. More importantly, although they conceptualised vocabulary as a three-factor model, they found that the best-fitting model showed vocabulary as comprised of only two dimensions: breadth and the combination of depth and fluency. Again, as in Kremmel et al.'s (2017) study, the two dimensions overlapped significantly ($r = .88$, $p < .001$), showing that they are very highly interrelated with each other, even in NS. Although the authors claim that these dimensions are different from each other, this extremely high correlation questions their independence as constructs. Therefore, their findings seem to suggest that vocabulary knowledge could be considered a unique dimension.

6.1.2.3. Vocabulary knowledge and other lexical aspects

Other SEM studies have employed this technique to test the relationship between vocabulary knowledge and other linguistic aspects related to lexical mastery, such as morphological awareness. Generally, these studies considered morphological awareness and vocabulary knowledge as independent, but related, constructs. For example, Zhang and Koda (2011) investigated the contribution of lexical inferencing and morphological awareness on L2 vocabulary knowledge by 130 adult EFL learners in China. Vocabulary knowledge was understood as knowledge of size (VLT) and depth (WAT), morphological awareness involved form recognition of the base of unknown derivatives, and lexical inferencing comprised meaning recognition of derivative words based on their suffixes. They found a strong relationship between all these aspects, with morphological awareness contributing to L2 vocabulary knowledge directly (explaining 19% of its variance) and indirectly through lexical inferencing, and lexical inferencing explaining about 15% of the variance in vocabulary knowledge.

The interrelation between vocabulary knowledge and morphological awareness has also been evidenced longitudinally. Kieffer and Lesaux (2012b) studied the concurrent development of these two types of linguistic knowledge on 90 Spanish children learners of English over three years. Conceptualising vocabulary as oral form–meaning link recognition (PPVT), and morphological awareness as base-form recognition of derivatives, they found a strong and statistical significant relationships (*r* =.68) between rates of growth in morphological awareness and vocabulary knowledge, after controlling for aspects such as word reading or phonological awareness. This suggests that when learners improve their knowledge of morphological awareness, they would also develop their vocabulary knowledge (i.e., size). This longitudinal relationship between vocabulary size and morphological awareness was found to be bidirectional by McBride-Chang et al. (2008). They studied the knowledge of morphological compounding and vocabulary size among 663 monolingual children of Cantonese, Mandarin and Korean at two points in time during one year. Their findings suggest that morphological knowledge in Time 1 significantly predicted vocabulary knowledge in Time 2, and, in turn, vocabulary knowledge in Time 1 predicted morphological knowledge in Time 2. Overall, these results show that as knowledge of one aspect develops, so does knowledge of the other.

These three studies seem to suggest that morphological and vocabulary knowledge are related, but independent, constructs. However, this claim has been contested in a more recent study by Spencer et al. (2015). Unlike in the previous research, they measured these concepts employing multiple tests and different types of response and administration of these tests (e.g., written, oral, multiple-choice or open answers). Morphological awareness involved nine different tasks tapping into various types of morphological knowledge (e.g., derivations, inflections and compounds), and vocabulary was assessed with a meaning-recall task measuring size, a task using words in context (grammatical functions, collocations and register) and an association task (synonyms and antonyms). An identical set of 23 words was assessed across all the aspects of knowledge to control for the differential knowledge of specific vocabulary items. In two different studies with over 180 English children, they found that a unidimensional model where vocabulary and morphological awareness comprised a unique construct provided the best fit to the data. The authors concluded that all the morphological and vocabulary tasks measure the same underlying ability, although they

vary in their difficulty, and that what one knows about a word affects all other types of knowledge of that word.

Overall, this review shows that some aspects of vocabulary knowledge have been investigated by means of SEM, although they have generally been studied marginally as part of a larger focus. That is, these studies do not typically examine the structure of vocabulary knowledge on its own, but with the aim to discover its contribution to reading comprehension and other linguistic skills (Babayiğit, 2015; Kieffer & Lesaux, 2012b; Kremmel et al., 2017; McBride-Chang et al., 2008; Spencer et al., 2015; Tannenbaum et al., 2006; Zhang, 2012; Zhang & Koda, 2011) or to examine the importance of language motivation and strategies to vocabulary acquisition (Ma, 2015; Tseng & Schmitt, 2008). As a consequence, their measures of vocabulary did not normally include individual, detailed tests of the different word knowledge components. Rather, they generally used a single instrument for vocabulary size (Babayiğit, 2015; Kieffer & Lesaux, 2012b; McBride-Chang et al., 2008), one test for depth (Kremmel et al., 2017), one measure for size and one for depth (Ma, 2015; Tseng & Schmitt, 2008; Zhang, 2012; Zhang & Koda, 2011), or employed multiple lexical measures but assessing skills in reading ability (e.g., Tannenbaum et al.'s (2006) test of timed production of sentences), rather than knowledge of the different vocabulary components. Therefore, these studies usually overlook the complex, multidimensional nature of vocabulary acquisition, and cannot inform of the relationships among word knowledge components or between the various components and overall vocabulary knowledge.

6.1.2.4. Analysis of the multidimensional structure of vocabulary knowledge

To the best of my knowledge, there are only two studies that have employed SEM for the primary purpose of empirically investigating the multifaceted nature of vocabulary knowledge and the relationships between its underlying components: Kieffer and Lesaux (2012a), and Scott, Flinspach, Vevea, and Castaneda (2017).

Kieffer and Lesaux (2012a) studied the English vocabulary knowledge of 584 sixth-grade students from diverse language backgrounds (including L1 and L2 English students). They employed seven reading-based measures assessing various word knowledge aspects. Using a multiple-group confirmatory factor analysis, they found

that vocabulary knowledge was comprised of three highly connected, but distinct, dimensions: breadth, contextual sensitivity and morphological awareness. This three-dimensional model had a better fit than a two-dimensional model (combining the first two factors in one) and a unidimensional model, and held both for L1 and L2 students separately. The authors concluded that a three-dimensional model was the best representation of vocabulary knowledge.

However, a closer inspection of the measures they employed raised some issues about their interpretation of this multidimensionality. Kieffer and Lesaux (2012a) had originally hypothesised that vocabulary knowledge was comprised of five aspects: breadth, depth, academic vocabulary, contextual sensitivity and morphological awareness, and tried to discover which of those aspects had empirical support as independent dimensions of vocabulary. The seven receptive tests used to assess those dimensions were:

- *General synonyms*. Multiple-choice test where students select a synonym of a target word.
- *Academic synonyms*. As the previous test, but with academic target words.
- *Semantic associations*. Students have to select three words, among six options, that are paradigmatic or syntagmatic associations of a target word.
- *Multiple meanings*. Students are given a sentence with a high-frequency target word on it and asked to select the sentence, among four options, in which the target word is used with the same meaning.
- *Context clues*. Multiple-choice task where students read a sentence with a rare (unknown) word in it and must select the word that has the same meaning as the target word, using the contextual information from the sentence.
- *Real word decomposition*. A word with a derivational suffix is given, and students need to extract the base of the word to fill a short sentence (e.g., *complexity* → *complex*, 'The problem is _____').
- *Nonword derivation*. Students need to complete a sentence (e.g., 'The man is a great _____') by choosing a nonword (e.g., *tranter*) with the

appropriate derivational suffix from among four options (*tranter, tranting, trantious, trantiful*).

In the final three-dimensional model proposed by Kieffer and Lesaux (2012a), the factor of vocabulary breadth comprised the academic synonyms, general synonyms, multiple meanings and semantic associations measures; the morphological awareness factor was represented by real word decomposition and nonword suffix choice; and the contextual sensitivity was measured by the context clues test. Vocabulary breadth was found to interrelate at the level of r =.85 with contextual sensitivity and r =.76 with morphological awareness, and the last two factors correlated at r =.71.

Four main limitations can be drawn from these findings. First, they investigate the multidimensionality of vocabulary based only on reading-based (thus receptive) measures. Therefore, other important vocabulary components are overlooked. Second, the tasks included as representations of vocabulary breadth are actually measures that lie under depth of vocabulary in most frameworks (i.e., knowledge of associations and polysemy) (Nation, 2013; Qian, 2002; Read, 1998). Thus, this factor is essentially combining breadth and depth of vocabulary knowledge. Third, the contextual sensitivity factor seems to reflect the ability to infer meaning from context, rather than a purely lexical aspect of vocabulary knowledge. Therefore, it came out as a different dimension in their study because it is representing more a reading skill than vocabulary knowledge. Finally, the morphological awareness factor measures relational and syntactical derivational knowledge, not distributional knowledge (see Chapter 2, Section 2.2.1.3). This might have led this factor to tap more into metalinguistic, morpho-syntactic skills than vocabulary knowledge, which in turn resulted in this aspect emerging as a distinct, independent dimension.

In conclusion, the model suggested by Kieffer and Lesaux (2012a) does not seem to be assessing the nature of only vocabulary knowledge, but also other linguistic abilities that can influence vocabulary. It seems that the factor that is in fact representing the vocabulary knowledge construct, as it is generally conceptualised in lexical studies, is just the one they called ‘breadth’. Hence, this questions the actual multidimensionality of the construct, and could indicate that vocabulary knowledge is indeed unidimensional. Therefore, although this is a sound study and provides a

relatively comprehensive investigation of the nature of vocabulary knowledge, these issues limit the conclusion of the multidimensionality of vocabulary.

The other study that examined the structure of vocabulary knowledge (Scott et al., 2017) addressed one of the limitations in Kieffer and Lesaux's (2012a) study, namely the conceptualisation of vocabulary, by assessing only word knowledge aspects within it. Scott et al. (2017) investigated the multidimensionality of vocabulary knowledge by testing 5,269 elementary school children on their knowledge of various vocabulary components for four target words. As in the previous study, the authors compared a unidimensional and a multidimensional model of vocabulary by means of confirmatory factor analysis, and found that the multidimensional model had a better fit. Thus, they concluded that their results provide empirical support for a multidimensional understanding of word knowledge. The six dimensions involved in this final model were:

- *Self-assessment*. Similar in format to the VKS, but without demanding any demonstration of knowledge of the target word.
- *Semantically related words*. Measures recognition of associations of a target word from four options.
- *Morphology*. Requires recognition, from three options, of the correct derivative word that relates to the target word.
- *Words used in context*. Requires selecting the most appropriate sentence from four options.
- *Definitions*. Meaning recognition test.
- *Parts of speech*. Recognition of the target word's word class.

There are two key reasons that make Scott et al.'s (2017) study more appropriate for the analysis of vocabulary knowledge than Kieffer and Lesaux's (2012a) study. First, as mentioned above, their measures include aspects of vocabulary knowledge that are described in most vocabulary frameworks (e.g., form–meaning recognition, word class recognition, recognition of associations). Second, they measure the same target words across all the vocabulary aspects. This not only represents the general word knowledge aspects better, but also ensures that the results in each measure are not an artefact of the ease or difficulty of the specific set of target words tested.

Nevertheless, there are still some features of the study design and analysis that make their results questionable. First of all, the comparison of the unidimensional and multidimensional models is based only on model fit indices. While these indices provide an indication of a better model fit, the information criteria indices (AIC and BIC) are the values that should be used to make the final decision on which is the best model. This information is missing in their study. Secondly, even though the original test they designed included a set of 24 target words assessed across the six dimensions, the authors only include the results of four of these words in the vocabulary knowledge models they measure. It is possible that the inclusion in the model of all the assessed words had yielded different results. What is more important, strong conclusions cannot be drawn with information retrieved from only four cherry-picked words. Thirdly, the measurement instruments used to assess the various types of knowledge in this study are slightly inconsistent, in that they combine self-assessment scales with demonstration of knowledge tests, and tasks such as 'words used in context' in which the type of knowledge measured is not clear (i.e., it could involve grammatical inaccuracy, grammatical functions, collocational knowledge, etc.). Finally, the authors only report unstandardised coefficients between the word knowledge variables and the vocabulary knowledge construct in both the uni- and multidimensional models. Reporting those coefficients instead of the standardised ones makes the results uninterpretable with the information provided. As a consequence, it is possible that the discriminant validity assumption is not satisfied. In other words, it can be that the model fit is good but the coefficients between the individual components and the vocabulary knowledge construct are suggesting lack of distinction between them as independent dimensions. All these issues mean that Scott et al.'s (2017) results regarding the multidimensionality of vocabulary knowledge are inconclusive.

Overall, although these two studies investigated vocabulary knowledge by assessing multiple of its aspects and provide the most complete examination of the construct to date, their limitations make it impossible to provide convincing evidence on the nature of vocabulary knowledge or the contribution of its different aspects to it. Therefore, there is still a need for a more comprehensive empirical investigation of whether these hypothesised dimensions of word knowledge can be distinguished reliably as unique constructs. This exploration would require following a more theoretically-grounded rationale for the selection of the types of knowledge under study,

assessing the same words across the different types of knowledge, and providing more detailed information for the comparison of different explanatory models.

The present study adopts this approach, and sets out to explore the overall nature of vocabulary knowledge as an independent construct by following the most widely accepted framework of word knowledge (Nation, 2013), concurrently measuring multiple vocabulary aspects both in recognition and recall, and providing empirical information about the relationships between components by means of SEM.

Before moving to the design of the present study, the following section introduces the structural equation modelling technique and the concepts and features to take into account for the design and administration of this methodology.

6.2. Structural equation modelling methodology

Structural Equation Modelling (henceforth SEM) is a set of statistical techniques that are theory-driven and follow a confirmatory approach to check the validity of a previously hypothesised model of relationships between variables (Byrne, 2016). SEM was adopted as the methodology for the present study because its specific characteristics (described in section 6.2.1) makes it the most appropriate method for the analysis of the relationships between vocabulary components, over other statistical techniques such as ANOVA's or multiple-regressions, and enable SEM to validate the multidimensional nature of the vocabulary knowledge construct.

The following sections will introduce the SEM technique and describe key concepts of its methodology. Then it will present the steps required to test a model of relationships. Finally, the different model fit criteria used in order to validate the model will be presented, as well as a brief discussion of which ones to use depending on the data in hand.

6.2.1. Introduction to SEM

SEM is a term used to refer to a set of related statistical procedures, and is also known as covariance structure analysis, causal modelling or path analysis with latent variables. SEM is considered a comprehensive analytical method because it comprises many other statistical techniques within it, such as multiple linear regression, factor

analysis, path analysis or ANOVA. This comprehensiveness means that SEM has some advantages that make it preferable over other conventional regression analyses.

First, SEM allows for the specification and analysis of both observable and unobservable variables, in contrast to regression analyses that can only analyse observed variables. These latent constructs are more reliable than any individual indicators, and less sensitive to the effects of the specific tasks used (Kieffer & Lesaux, 2012a). Hence, this allows us to explore the unobservable vocabulary knowledge construct by measuring its different components individually. Second, the multiple hypothesised relationships among variables specified in the model can be tested simultaneously instead of one by one. This means that the strength of each hypothesised link is analysed while taking into account all the other connections, allowing to estimate how closely the data fits the hypothesised model and reducing the chances of Type I error. This way, we can concurrently explore the interrelationships between the word knowledge components with each other and in relation to the latent vocabulary construct. Third, while regression models require the specification of independent and dependent variables, the variables in SEM can be dependent and independent at the same time. Fourth, SEM can analyse multivariate relations (e.g., when correlations are only expected among certain variables), direct effects (e.g., an independent variable affects another one directly) and/or indirect effects (e.g., when an independent variable affects another one through a mediating variable) (Byrne, 2016). This allows us to explore different types of relationships between word knowledge components. Fifth, the measurement error of variables can be identified, estimated and removed, allowing the researcher to interpret the relationships among factors and between multiple indicators and factors free of measurement error (Ullman, 2013). Finally, SEM applies a confirmatory, theory-driven approach to the data, in contrast to traditional analysis which are data driven (e.g., Exploratory Factor Analysis or Mokken analysis). This means that SEM can, and is used, to test the plausibility of theoretical models that explain the interrelations among a set of variables (Hu & Bentler, 1999).

For all these reasons, SEM is the preferred methodology for theoretical model building and testing, and has been commonly used in the behavioural and social sciences to represent relationships of dependency among variables and establish or validate theories (McDonald & Ho, 2002). I believe the time is ripe to apply this powerful technique to the analysis of overall vocabulary knowledge in an attempt to test

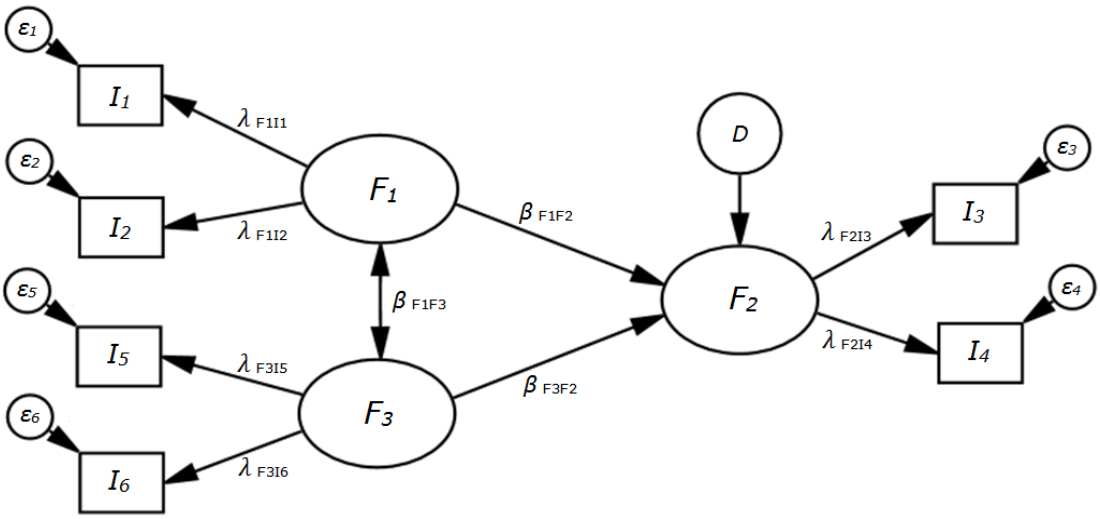
the nature of the vocabulary knowledge construct and set the basis for the development of a general theory of vocabulary acquisition.

Given the confirmatory, hypothesis-testing nature of this technique, the first step when applying SEM is to specify a hypothesised model of multiple relationships between variables, which is generally based on previous literature or researcher's intuitions (i.e., theory-driven). This model is then submitted to be empirically tested and confirm whether the model is supported by the data. As stated by Ullman (2013), the hypothesised model creates an estimated population covariance matrix, whereas the set of empirically collected data produces a sample (observed) covariance matrix. The purpose is to examine the extent to which the proposed estimated population covariance can fit into the sample covariance matrix. The closer the two matrices, the better the hypothesised model.

6.2.2. Basic concepts of SEM

In this section, important concepts and conventions employed in SEM will be introduced. In order to assist the explanation, Figure 13 provides a graphic example of a structural equation model.

Figure 13 Example of a structural equation model.



Note: *F* = factors or latent variables, *I* = indicator, *ε* = measurement error, *β* = structural coefficients, *λ* = factor loadings, *D* = disturbances.

6.2.2.1. Types of relationships

In a SEM model, the relationships between variables are specified by arrows. Two main types of relationships can be identified: correlations and causal relationships. The correlations between variables are represented by double-headed arrows (e.g., F_1 ←→ F_3), and they mean that the variables are interrelated but there is no hypothesised causal relationship between them. The causal or direct relationships between two variables are represented by single-headed arrows (e.g., F_1 → F_2), and show a directional effect where one variable influences the other, but not the other way around. Absence of an arrow between variables means that there is no hypothesised relationship between them. Absence of a single-headed arrow implies lack of direct cause-effect relationship between two variables, whereas absence of double-headed arrows implies that the variables have no common cause that has been omitted in the model (McDonald & Ho, 2002).

6.2.2.2. Nature and types of variables

SEM models comprise both observed and unobserved variables. The variables represented by rectangles (i.e., I_1, I_2, I_3, I_4 I_5 and I_6) are the *observed variables*, a.k.a. *measured variables, manifest variables* or *indicators*. They are connected to the factor they represent by single-headed arrows. The multiple indicators of a factor should be correlated with each other if they are under a common construct. The ovals (i.e., F_1, F_2 and F_3) generally represent the *factors*, also called *unobservable variables*, *latent variables* or *constructs*. They symbolise abstract, unobservable concepts that cannot be measured directly, but indirectly by means of their indicators. For example, in Figure 13 the factor F_1 is measured by its indicators I_1 and I_2. Factors are free of measurement error, which makes them propitious for the validation of hypotheses/theories.

Unlike more traditional multivariate techniques, SEM does not require the specification of dependent and independent variables. Instead, we speak of *endogenous* and *exogenous* variables. Endogenous variables ('of internal origin') are the effect of other variables; that is, they have at least one arrow pointing to them (like dependent variables). In Figure 13, indicators I_1 to I_6 and factor F_2 are endogenous variables. In the first case, the indicators are predicted by the particular factor they represent, and in the second case, factor F_2 is predicted by other factors (F_1 and F_3). Exogenous variables

('of external origin') are those which are not predicted by any variable, but explain or predict others. These predictor variables (i.e., F_1 and F_3) have arrows coming from it but no arrow pointing to it (independent variables). A particularity of SEM is that endogenous variables can also predict other variables (i.e., have arrows pointing in and out). Therefore, they can be dependent and independent variables at the same time. For example, F_2 is predicted by F_1 and F_3 (dependent), but it also predicts I_3 and I_4 (independent). This is one of its advantages over other techniques.

As in any method, dependent variables can never be perfectly predicted by the independent variables, and thus there are always some *measurement errors* or *disturbances* (residual errors) associated with these variables. This disturbance represent all omitted causes and any measurement error (i.e., variance that predictors did not predict). By convention, the measurement errors or disturbances will have an arrow pointing to the endogenous variable. However, this arrow does not indicate a direct relationship between them, but the source of measurement error. In Figure 13, each indicator has a measurement error, represented by the symbol ε. The variance of a measurement error means the unexplained variance of an indicator by the factor. For example, the variance of ε_1 is the variance of I_1 that cannot be explained by F_1. Similarly, the variance of a disturbance or residual error associated with a factor means the unexplained variance of that factor by the overall model. For example, the variance of *D* associated with F_2 is the variance of F_2 that cannot be explained by the model. The measurement errors of two indicators can be correlated. This suggests that the two indicators measure something in common that is not explicitly represented in the model (Kline, 2016). Therefore, a measurement error correlation reflects the assumption that the indicators are different but also interrelated to a certain extent.

6.2.2.3. Strength of relationship

There are two types of coefficients representing the strength of relationships between variables, depending on the nature of this relationship: factor loadings or structural coefficients. The symbols λ_{F1I1}, λ_{F2I1}, etc., denote *factor loadings*, and represent the strength of relationships between indicators and factors. For its part, the symbols β_{F1F2}, β_{F1F3}, and β_{F3F2} are called *structural* or *path coefficients*, and represent the strength of relationships between factors. When the strength of relationship between

factors is standardised[16], they are equivalent to standardised regression coefficients (β), and thus represent the direct effect from a predictor on the endogenous (dependent) variable. Therefore, the square value of β_{F1F2} ($(\beta_{F1F2})^2$) and β_{F3F2} ($(\beta_{F3F2})^2$) represent the proportion of variance of F_2 that can be explained by the predictors F_1 and F_3 respectively. The square value of the factor loadings (e.g., $(\lambda_{F1I1})^2$) represents the reliability indices for the particular indicator.

The elements that indicate the nature and size of the relationships between two variables in the population are called *parameters*. In SEM, a parameter is estimated for each relationship between variables, whether they are regressions or correlations. Parameters can be specified in the model as *fixed*, that is, set equal to some constant, or *free*, that is, to be estimated from the data. When a the model is too complex to be estimated, some parameters need to be set fixed.

6.2.2.4. Types of models in SEM

A model is a statement about relationships between variables. In the development of any SEM representation two types of models are generally specified: the *measurement model* and the *structural model*. The measurement model is also called Confirmatory Factor Analysis (CFA), and plays a very important role in model building because CFA can be used to confirm how well a latent variable can predict or reflect its representative indicators. In other words, check whether the indicators used to measure a particular factor belong to such factor. This model is comprised of a latent variable and its several indicators. For example, in Figure 13 a measurement model would be the section comprised of F_1, I_1, I_2, its measurement errors and its factor loadings. The structural model is the combination of: 1) the hypothesised structural relationships among a set of latent variables (this is known as the *path model*), and 2) the individual measurement models. In Figure 13, the section comprising F_1, F_2, F_3, the disturbance and the structural coefficients would be a path model. When validating these models, the measurement model can be tested individually, while the structural model requires the combination of various measurement models.

[16] The structural coefficients reported in this study are all standardised.

6.2.3. Steps in SEM

When building SEM models, five basic steps need to be carried out: 1) model specification based on theory, 2) model identification, 3) data and parameter estimation, 4) assessment of a model's goodness-of-fit, and 5) model respecification, if applicable (Bollen & Long, 1993; McDonald & Ho, 2002). The following sections will briefly describe these steps.

6.2.3.1. Model specification

Due to the theory-driven nature of SEM, the model specification step requires the development of a hypothesised model based on previous findings or theories in the area. As mentioned above, two types of models need to be specified in SEM, the measurement model and the structural model. In the measurement model we need to represent the set of observed variables as the indicators of a particular latent variable. It is generally represented as a CFA model, in which the indicators load on their particular latent variables, which are in turn all intercorrelated. Within the structural model, we illustrate the hypothesised relationships between the latent variables, which form the path model. This path model reflects the correlation (or covariance) matrix between the latent variables, and involves the specification of direct and/or indirect effects between them. The combination of the measurement and path models comprise the structural model (McDonald & Ho, 2002).

Given that SEM is an a priori approach, when specifying and reporting SEM models the researcher should provide theoretical explanations of the relationships reflected in the model, as well as describing the absence of other relationships (Hoyle & Panter, 1995).

6.2.3.2. Model identification

The first requirement for a model to be considered valid is for it to be identified. Identification of a model implies that the available data is sufficient to solve the equations, and thus estimate the unknown parameters in the models. If the model is unidentified, then it cannot be solved because there are more unknown than known variables. To avoid unidentification of a model one cannot include all the possible

relationships in it. That is, some restriction or constraints are required. Therefore, this model identification step involves analysing whether the model can produce a unique value for each parameter in the model.

Moreover, for the measurement model to be identified, each indicator can load only on one latent variable, and each latent variable needs to be comprised of at least two indicators if the latent variables are correlated, and at least three if they are not. For the path model to be identified, the disturbances of endogenous latent variables need to be uncorrelated (McDonald & Ho, 2002).

6.2.3.3. Data and parameter estimation

The data and parameter estimation stage implies obtaining an estimated coefficient for each unknown parameter in the model. Multiple estimation methods exist. The most commonly employed are the Maximum Likelihood (ML) and the Generalised Least Squares methods, which are based on the assumption of multivariate normal data. Other methods such as the robust maximum likelihood (MLR) and various weighted least squares (WLS) (e.g., weighted least squares means and variance adjusted (WLSMV)) do not assume multivariate normality, and thus are less likely to reject valid models under non-normality conditions (Hu & Bentler, 1999). Hence, although non-normal data can affect the chi-square and fit indices, this can be handled by employing these robust estimation methods. Since in the social sciences the data is generally non-normally distributed (Field, 2013; Micceri, 1989; Ramzan, Zahid, & Ramzan, 2013), the MLR or WLSMV estimation methods are considered to be more appropriate (Finney & DiStefano, 2013).

ML estimation methods have been found to be the most often used ones in language learning SEM studies (In'nami & Koizumi, 2011). That is because ML methods are better at estimating parameters with continuous variables, and thus are preferable for this type of variables. If the continuous variables are nonnormally distributed, the MLR estimation is recommended (Finney & DiStefano, 2013). Thus, this will be the estimation method employed in my study for the continuous variables.

6.2.3.4. Model fit

Since the hypothesised relationships between variables in a model have to be specified a priori, assessing the goodness-of-fit of the model is an essential stage in SEM. This model fit step tests the degree to which the hypothesised model fits the data. The better the model fit is, the closer the estimated population matrix and the observed matrix are, and thus, the better the hypothesised model is. This stage requires the inspection of various types of model fit indices, which are described in detail in Section 6.2.4.

6.2.3.5. Model respecification

After having tested the hypothesised model, the model respecification stage studies the ways to improve the model-data fit. This can be done by deleting statistically nonsignificant paths or by adding new paths. This respecification of the original model, however, must always be supported by the theory, and not be driven by model-fit improvement. Moreover, since SEM is a confirmatory technique, this model respecification step should not include many or very dramatic changes (McDonald & Ho, 2002).

6.2.4. Model evaluation

SEM examines the extent to which the estimated population matrix of relationships described in the hypothesised model is consistent with the observed matrix of relationships included in the dataset. Model evaluation analyses how close the two matrices are. The more similar they are, the better the hypothesised model is. As such, the model evaluation is an indication of how well the model that best represents the data reflects the underlying theory (Hooper, Coughlan, & Mullen, 2008). Hence, SEM assesses whether a theoretical model can be rejected by the data through evaluating its fitness (Arbuckle, 2016).

In order to examine how well the hypothesised model fits the data, a comprehensive list of global model fit measures has been proposed. These global goodness-of-fit indices can be classified in three categories: 1) Absolute fit indices, 2) Comparative fit indices, 3) Parsimony fit indices (Hooper et al., 2008). Each of these

types of indices provide information about a certain feature of the model, and thus they are all important. Since multiple indices have been proposed for each of these categories, the following sections will focus on describing those that will be reported in this study.

6.2.4.1. Absolute fit indices

Absolute fit indices determine how well the a priori hypothesised model reflects the sample data. That is, they are an indication of how well the proposed theory fits the data (Hooper et al., 2008). They also show which proposed model has the superior fit (McDonald & Ho, 2002). Unlike other indices, they are not based on the comparison with a baseline model, but measures how well the model fits in comparison with a null model. Included in this category are the Chi-Squared test, root mean square error of approximation (RMSEA), the standardised root mean squared residual (SRMR), the Goodness-of-fit statistics (GFI) and its adjusted version (AGFI). The last two indices have been found to perform worse than the other measures, and thus are not recommended for model evaluation (Hu & Bentler, 1999).

- **Model chi-square (χ^2).**

The Chi-Square value is the traditional measure for evaluating overall model fit. It examines the amount of discrepancy between the observed and the estimated covariance matrices (Hu & Bentler, 1999). In SEM, a good model fit would be reflected by a non-significant chi-square ($p \geq .05$). This is because a significant chi-square would suggest that the estimated covariance matrix is significantly different from the observed covariance matrix, which is undesirable. A non-significant finding suggests that the two matrices are close, implying acceptance of the hypothesised model (Barrett, 2007).

Although the chi-square test is a very popular model fit index, there are a number of limitations that affect its use (Kline, 2016). First, it assumes multivariate normality of data, and thus deviations from normality may result in model rejection (i.e., $p < .05$) even when the model is valid (Hooper et al., 2008). Secondly, chi-square is sensitive to sample size, which translates in rejection of models under large-sample conditions, and lack of discrimination between good and bad fitting models under small-sample conditions (Cheung & Rensvold, 2002). In order to minimise the impact of sample size on the chi-square value, reporting the normed chi-square (χ^2/df) has been recommended,

and it has been suggested that this value should be between 1 and 3 (Hair, Black, Babin, & Anderson, 2010).

The chi-square statistics should be reported in any SEM model despite its limitations, along with its degrees of freedom (*df*) and *p* value (Kline, 2016). However, a model should not be rejected on the sole basis that the chi-square is significant, provided that the other fit indices (i.e., RMSEA, SRMR) are good (Kenny, Kaniskan, & McCoach, 2014).

- **Root mean square error of approximation (RMSEA)**

The RMSEA represents how well the observed covariance matrix approaches the estimated population covariance matrix (Byrne, 2016). The RMSEA favours parsimony, that is, has a better value when the model has less number of parameters (i.e., is simpler). RMSEA of <.05 or <.06 is considered good fit, and <.08 is considered an acceptable fit (Bagozzi & Yi, 1988; Hu & Bentler, 1999). The RMSEA should be reported in conjunction with its *p* value (good fitting model showing *p*>.05) and its confidence interval (low limit around 0 and upper limit generally below .08).

- **Standardised root mean square residual (SRMR).**

The SRMR is the square root of the average difference between the residuals in the sample covariance matrix and in the hypothesised covariance matrix. Well-fitting models should obtain values between 0 and .05, with 0 indicating perfect fit (Hooper et al., 2008), although models are still considered acceptable if SRMR <.08 (Hu & Bentler, 1999). An advantage of SRMR over RMSEA and CFI is that its calculations are not based on the chi-square, and thus this measure is not affected by the chi-square's limitations. For this reason, it is recommended that SRMR always be reported (In'nami & Koizumi, 2011).

6.2.4.2. Comparative fit indices

Comparative fit indices, also called incremental, assess the relative improvement in model fit by comparing the hypothesised model with a baseline (null) model in which the observed variables are uncorrelated (McDonald & Ho, 2002). Some examples are the Normed-fit index (NFI), the Tucker-Lewis index (TLI), or the

Comparative Fit Index (CFI). Due to limitations of the other incremental fit indices (Hu & Bentler, 1999), the most commonly reported measure is the CFI.

- **Comparative fit index (CFI)**

The CFI measures the model by comparing the chi-square value of the hypothesised model to the chi-square of the null model (Byrne, 2016). The CFI represents the percentage of lack of fit between the null model to the researcher's model. Therefore, the higher the CFI, the better the hypothesised model. The values range from 0 to 1. Although traditionally a value greater than .90 was accepted, a CFI ≥.95 is presently required as an indicative of good fit (Hu & Bentler, 1999). This index is not largely affected by sample size, performing well in small-sample conditions (Hooper et al., 2008).

6.2.4.3. Parsimony fit indices

In any regression analysis (i.e., the basis of SEM), the parsimony principle holds. This principle refers to the idea that the simplest explanation about the greatest number of relationships is preferred over more complex ones. That is, simpler models (i.e., with less freely estimated parameters) are favoured. The information criteria indices are the ones employed to compare the parsimony of multiple models.

- **Information criteria indices**

The most common information criteria indices are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). These indices estimate the likelihood of the model to fit in hypothetical, similar-size samples drawn from the same tested population. These measures are used when comparing multiple models estimated with the same data. They indicate which of the competing models is the most parsimonious (i.e., simplest), and thus the best-fitting model. Smaller AIC or BIC values suggest a better fit.

Table 42 summarises the cut-off thresholds for the model fit indices relevant for this chapter.

Table 42 Summary of fit indices and their threshold levels.

Fit index	Acceptable Cut-off criteria	Description
	Absolute Fit Indices	
Chi-square (χ^2)	Low χ^2 relative to degrees of freedom with $p > .05$	
Normed chi-square (χ^2/df)	1-3	Adjusts for sample size
RMSEA	<.05 or <.08	Favours parsimony. Value less than .03 represents excellent fit (Hooper et al., 2008).
SRMR	Less than .08	Good models have small SRMR (close to 0).
	Incremental Fit Indices	
CFI	Values greater than .95	Normed. 0-1 range.

The aforementioned global goodness-of-fit indices are the most commonly reported measures of model evaluation, and they are used to decide whether a model should be rejected. Nevertheless, there are other aspects of the internal structure of a model that need to be considered when deciding how well a model fits the data. These involve checking the unidimensionality, validity and reliability of models.

Unidimensionality. Unidimensionality is an essential feature of a latent variable that refers to the relationships between indicators and factors (Awang, 2012). Unidimensionality of a measurement model is achieved when all factor loadings are acceptable and positive. An acceptable factor loading should range between .40 (Hooper et al., 2008) and .95 (Bagozzi & Yi, 1988). Factor loadings <.40 indicate a low relationship between the indicator and the factor it belongs to, suggesting that it is less reliable and less representative of the factor. Conversely, a factor loading >.95 indicate a very high relationship between an indicator and the factor, suggesting that the indicator is redundant. These cases are recommended to be deleted to achieve better unidimensionality.

Validity. In order to ensure that the measurement instruments assess what they are supposed to assess for each latent construct three types of validity tests should be performed.

- *Convergent validity*. It refers to how well a factor is measured by its indicators. It is achieved when all items in a measurement model are statistically significant, and also by obtaining an average variance extracted ≥ .50 (see Reliability section below). Retaining the low factor loading items could cause the model to fail to achieve convergent validity.
- *Construct validity*. This type of validity is attained when the global fit indices achieve the required threshold levels.
- *Discriminant validity*. It refers to the relationships between latent variables, and indicates whether the model is free from redundant items. Structural coefficients between exogenous latent variables should not exceed .85 (Kenny, 2016). If they are ≥.85, it indicates that there is a lack of discriminant validity within the latent variables in the model. In other words, the latent variables cannot be differentiated from each other, and thus the constructs are empirically redundant (Awang, 2012).

Reliability. The reliability of the measurement model in measuring the intended latent construct was tested in several ways.

- *Internal consistency*. It refers to how strongly the various indicators of a factor hold in measuring a construct. It was assessed by means of Cronbach's alpha, with α ≥.70 considered reliable.
- *Internal item reliability*. It refers to the reliability of each indicator in a measurement model. It is computed by squaring the factor loadings, and it represents the proportion of variance in the indicators that can be explained by the factor. The internal item reliability should be ≥.20 (i.e., factor loadings ≥.40). The unexplained percentage of variance in an indicator is the measurement error, calculated by 1-squared loading.
- *Composite Reliability (CR).* It assesses the reliability and internal consistency of all the indicators of a factor, and thus, of the factor as a whole. By convention, CR should be >.60 (Awang, 2012). Given to Cronbach's alpha limitations (i.e., assumption that all indicators are

equally reliable), the CR is assumed more appropriate to measure internal consistency (Hair, Hult, Ringle, & Sarstedt, 2017). This coefficient is achieved by computing the following formula:

$$CR = \frac{(sum\ of\ standardized\ factor\ loadings)^{2}}{((sum\ of\ standardized\ loadings)^{2} + (sum\ of\ 1 - \ squared\ factor\ loadings))}$$

– *Average Variance Extracted (AVE)*. AVE indicates the average percentage of variation explained by all the indicators of a factor. It should be ≥.50. The formula is:

$$AVE = \frac{sum\ of\ squared\ standardized\ factor\ loadings}{number\ of\ items\ in\ a\ model}$$

It is important to conclude this section by reminding that a SEM model should always be examined based on the theory. The researcher's knowledge of the theory should lead the research process, employing the fit indices as guides.

6.2.4.4. Reporting fit indices

The abundance of fit indices available has led to a lack of consensus regarding which model fit indices to report and their cut-off points. It is agreed upon that it is not necessary (or even realistic) to report all the possible fit indices (Hooper et al., 2008; McDonald & Ho, 2002). It is not good practice either to provide only one model fit measure or select those that show the best model fit. In order to solve this problem, different authors have suggested various combinations of fit indices to be reported in SEM analysis. These combinations generally include at least one index from the different categories, since each type has different sensitivities to model features such as sample size, model complexity or normality of the data (Fan & Sivo, 2007; Hair et al. 2010). Therefore, the selection of the specific indices is not arbitrary, and needs to be based on the particular conditions of the data and the sample (Hu & Bentler, 1999). Table 43 summarises the suggestions by different authors on which fit indices to report.

Table 43 Summary of fit indices combinations.

Authors	**Fit indices**										
	χ^2	*df*	χ^2 *p* value	RMSEA	RMSEA +CI	CFI	SRMR	R^2 of each equation	TLI	AIC	GFI
Kline (2016)	✓	✓	✓		✓	✓	✓				
Boomsma (2000)	✓	✓	✓	✓ (or)	✓ (or)		✓	✓			
Hu & Bentler (1999)				✓ (or)		✓ (or)	✓		✓ (or)		
Hooper et al. (2008)	✓	✓	✓		✓	✓	✓			✓	
In'nami & Koizumi (2011)	✓	✓	✓	✓ (or)		✓ (or)	✓		✓ (or)		
Russell (2002)				✓ (or)		✓ (or)	✓		✓ (or)		
Kashy, et al. (2009)	✓			✓		✓ (or)			✓ (or)		
Widaman (2010)	✓			✓		✓			✓		
Bandalos & Finney (2010)	✓			✓		✓	✓		✓		
Mueller & Hancock (2010)				✓		✓ (or)	✓		✓ (or)		
Shiotsu & Weir (2007)	✓	✓	✓	✓							✓
McDonald & Ho (2002)	✓	✓		✓			✓				✓

Using extensive simulation studies, Hu and Bentler (1999) reviewed the best model-fit combinations to be reported, and recommend a two-index report including always the SRMR complemented by either TLI, RMSEA or CFI. They recommend the combination of CFI and SRMR when sample size is less than 250, as TLI and RMSEA tend to over reject valid models under this sample size. However, most authors seem to recommend reporting more than two model indices (In'nami & Koizumi, 2011).

Following the guidelines from previous studies and the characteristics of my data (relatively small sample size and multivariate non-normality), I will report the chi-square test statistic, including the number of degrees of freedom and its significance level, the RMSEA with the 90% Confidence Interval and significance level, the SRMR as a feature of the residuals and the CFI. These have been found to be the most insensitive combination to sample size, model misspecification and parameter estimates (Hooper et al., 2008).

6.3. Purpose and research questions

This study aims at investigating the structure of the vocabulary knowledge construct in L2 learners of English, operationalised as knowledge of the form–meaning link, derivatives, multiple meanings and collocations. The main objective is to explore the overall nature of vocabulary knowledge and obtain empirical confirmation of its multidimensionality.

The following research questions will be addressed:

1. How strongly do the various word knowledge components contribute to the vocabulary knowledge construct? Which component has the strongest effect in predicting vocabulary knowledge?
2. Can predictive relationships be described between certain components?
3. What is the best way of conceptualising the relationships between various word knowledge components?

The next section describes the aspects of the research design implemented to answer these questions.

6.4. Methodology

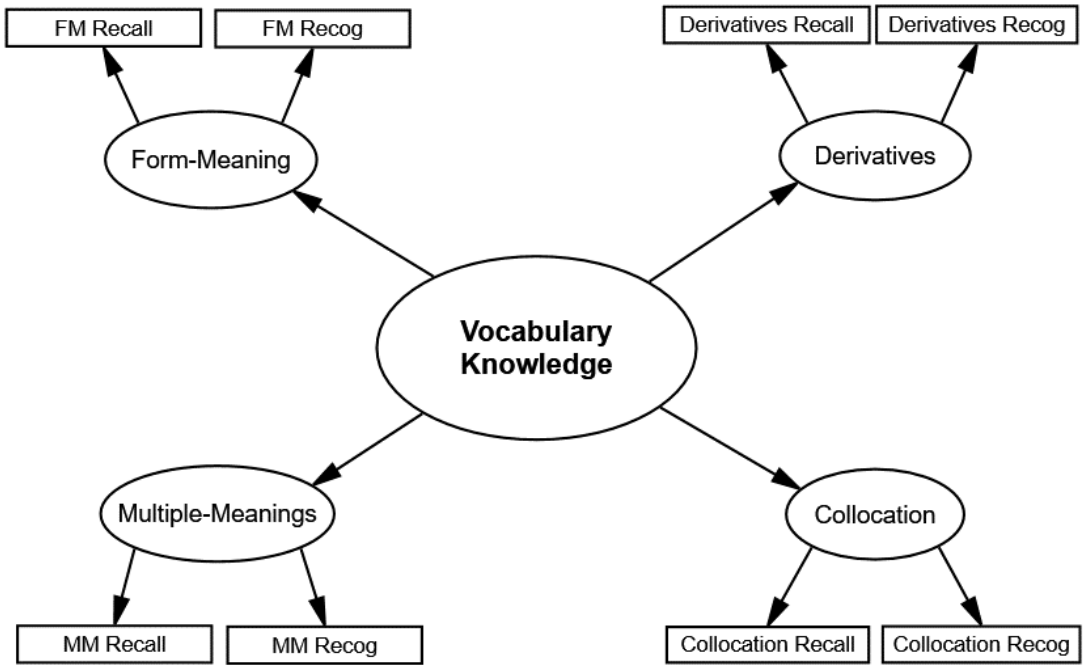
The present study empirically examines the relationships among word knowledge components and tries to validate the multidimensionality of vocabulary knowledge using latent variables. An advantage of using latent variables is that the relationships among the variables are examined free of measurement error and that the reliability of the measures does not affect the relationships (Tannenbaum et al., 2006). Thus, this analysis can provide an accurate representation of the relationships between the aspects of word knowledge.

The study adopts the research design described in Chapter 3. The dataset employed in the analysis was the one obtained for Study 1 (Chapter 4), and therefore the participants are those presented in Chapter 4.

6.4.1. The hypothesised model

As is required in SEM, the model to be tested was specified prior to the design of the measurement instruments and data collection. Thus, a model of general vocabulary was hypothesised based on previous research. Although there is no generally accepted theory of vocabulary acquisition, vocabulary is commonly described as a multifaceted construct in nature, and, as such, would involve the acquisition of various dimensions. The most widely accepted conceptualisation of vocabulary knowledge is Nation's (2013) framework of what is involved in knowing a word. According to this framework, vocabulary knowledge is seen as comprised of multiple word knowledge components, with each of these components being made up of both recognition and recall knowledge. I modelled this conceptualisation as a structural equation diagram as seen in Figure 14. This hypothesised model represents vocabulary as a general, underlying latent construct comprised of smaller sub-constructs (i.e., form–meaning link, derivatives, multiple meanings and collocations) which, in turn, are represented by their recognition [Recog] and recall [Recall] masteries.

Figure 14 Hypothesised model of relationships between word knowledge components.



Based on Nation's description of vocabulary knowledge, the hypothesised relationships between components are expected to be as follows. As for the relationships in the path model (i.e., between the latent variables), vocabulary knowledge involves mastering all the components, and therefore, it is anticipated that all these components will represent vocabulary very strongly. Much previous research has suggested strong interrelationships between the aspects (e.g., Milton & Fitzpatrick, 2014; Schmitt, 1998, 2014). This means that high structural coefficients are expected to be found between vocabulary knowledge and the four latent components.

In addition, the form–meaning link has been claimed to be the essential aspect for vocabulary knowledge (e.g., Agustín Llach & Moreno Espinosa, 2014; Laufer & Goldstein, 2004). Therefore, a stronger relationship between vocabulary knowledge and form–meaning link is likely. Moreover, collocational knowledge has been found to pose exceptional problems to L2 learners (Bahns & Eldaw, 1993; Granger, 1998; Gyllstad & Wolter, 2015; Peters, 2016; Webb et al., 2013). As such, it is probable that its relationship to overall vocabulary knowledge is lower than that of other components. However, all these previous relationships between components have been found in studies that analyse the links between components independently. Given that SEM

examines the hypothesised relationships simultaneously, it may be that interesting results will be yielded.

Regarding the measurement models for each latent variable, each word knowledge factor included two indicators, recognition and recall mastery. The recognition and recall knowledge of a component are considered to be correlated (e.g., Laufer & Goldstein, 2004; Milton & Hopkins, 2006). Consequently, I formulate that the recognition and recall indicators of each component will load highly into their specific factor, suggesting that they indeed belong to the factor they represent and that they are interrelated through that factor. Finally, it would be interesting to check which of the indicators, recognition or recall, is a better representation of a factor. For example, given that receptive knowledge of a component generally exceeds productive knowledge (Webb, 2008), it may be expected that the recognition mastery has a stronger bond to a particular component.

By using SEM, I seek to examine the relationships between components specified in the model, as well as empirically test the validity of the multidimensional conceptualisation of vocabulary knowledge.

6.4.2. Data analysis

In the first instance, some preliminary analyses were conducted. Firstly, the data was checked for missing values and outliers, since SEM is sensitive to them. Data screening confirmed that there were no outliers or missing data in the variables relevant to the study. Secondly, the normality of the data was tested, as SEM assumes univariate and multivariate normality. The univariate normality of the variables included in the model was tested in Study 1 (Chapter 4), and they showed to be non-normally distributed. Then, the multivariate normality of the data was assessed by means of Mardia's Multivariate Normality Test (Mardia, 1970) in the R-software with the package *MVN* (version 4.0.2) (Korkmaz, Goksuluk, & Zararsiz, 2016). Mardia's Test suggested that the data was multivariate non-normal (i.e., multivariate skewness and kurtosis $p < .05$), as expected. Therefore, given that multivariate non-normal data can affect the chi-square and fit indices, the robust estimation methods were employed for the estimation of parameters to account for this non-normality. MLR method was applied for continuous variables, and WLSMV for categorical ones.

An adequate sample size is necessary when using SEM, since parameter estimates and some fit indices are calculated based on it. Although there is no rule-of-thumb about the necessary sample size, most previous studies seem to agree that 100 to 150 participants is the minimum number adequate for analysis (Ding, Velicer, & Harlow, 1995). For this reason, a sample of around 100 participants was targeted. The sample size in this study (N = 144) can be considered small to medium (Kline, 2016).

The data was analysed using the R-Studio software (R Core Team, 2017), version 3.4.1. Structural equation models were fitted using the *lavaan* package, version 0.5-23.1097 (Rosseel, 2012).

Exploratory Factor Analysis was considered inappropriate for the purpose of this study since the factors and indicators are based on the extensive literature of the area and I wanted to confirm their belonging to the construct. For the purpose of this study, confirmatory factor analysis (CFA) is more appropriate than EFA, since, unlike the latter, CFA is a theory-driven method in which models can be tested using robust model fit indices and sophisticated statistical tests to determine which has a better fit to the data (Kieffer & Lesaux, 2012a). Therefore, I started by applying CFA. CFA is a special form of factor analysis employed to test whether the measures of a construct are consistent with the researchers' understanding of the nature of that construct (Awang, 2012). It is considered good practice to assess the fit of each latent construct and its indicators individually by means of CFA before testing the hypothesised structural model. This process is used to validate each latent variable and ensure that it represents a valid psychometric construct. If particular items are found to be weak (r <.40 and r^2 less than .20) they should be removed from the analysis (Hooper et al., 2008). The unidimensionality, validity and reliability of the models are also reported.

SEM was then used to explore the nature of overall vocabulary knowledge by assessing the hypothesised model. My instruments include two indicators per component (i.e., recognition and recall), satisfying the SEM requirement of a minimum of two indicators per factor for a measurement model to be identified. Because there is no generally accepted theory of how vocabulary is acquired, other causal models based on different traditions of understanding vocabulary knowledge were tested employing SEM.

The acceptability of the models was judged on the basis of the fit indices and threshold guidelines most commonly employed (e.g., Brown, 2015; Hu & Bentler,

1999; In'nami & Koizumi, 2011; Kline, 2016). A good model fit was indicated by a nonsignificant χ^2, χ^2/df of between 1 and 3, CFI of >.95, SRMS <.08, RMSEA of < .05 for good fit and < .08 for acceptable fit.

6.5. Results and interpretation

6.5.1. Descriptive statistics and correlations

Table 44 shows a summary of the participants' performance in each test in percentage scores and the correlation coefficients between the measures. Regarding the descriptive statistics, it can be seen that, on average, learners knew about half of the items for each test, and that the recognition mastery of an aspect always shows higher scores than its recall mastery. The wide gap between the minimum and maximum scores shows that participants had a wide range of proficiency levels. As for the interrelationships between components, Spearman's correlations were computed due to the non-normality of the variables. The correlations between word knowledge components range between .68 and .95, showing a high degree of relationship. This supports Schmitt's (2014) conclusion that word knowledge components seem to be highly interconnected.

Table 44 Descriptive statistics (%) and intercorrelations between components (Spanish group).

	FM Recall	FM Recog	Deriv Recall	Deriv Recog	MM Recall	MM Recog	Collo Recall	Collo Recog
FM Recall								
FM Recog	.773**							
Deriv Recall	.811**	.803**						
Deriv Recog	.786**	.828**	.945**					
MM Recall	.747**	.774**	.845**	.818**				
MM Recog	.720**	.766**	.835**	.831**	.848**			
Collo Recall	.772**	.749**	.827**	.812**	.817**	.771**		
Collo Recog	.678**	.700**	.763**	.768**	.752**	.737**	.806**	
Min.	10.00	40.00	5.00	18.75	6.67	31.67	15.00	25.00
Max.	95.00	100	91.25	97.50	90.00	96.67	95.00	100
M	53.61	82.81	51.64	61.08	49.69	70.76	60.66	79.69
SD	18.91	13.79	20.86	19.90	16.12	15.18	19.05	17.39

Spearman: $^{**}p < .01$, $N = 144$

FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition

6.5.2. Structural equation modelling analyses

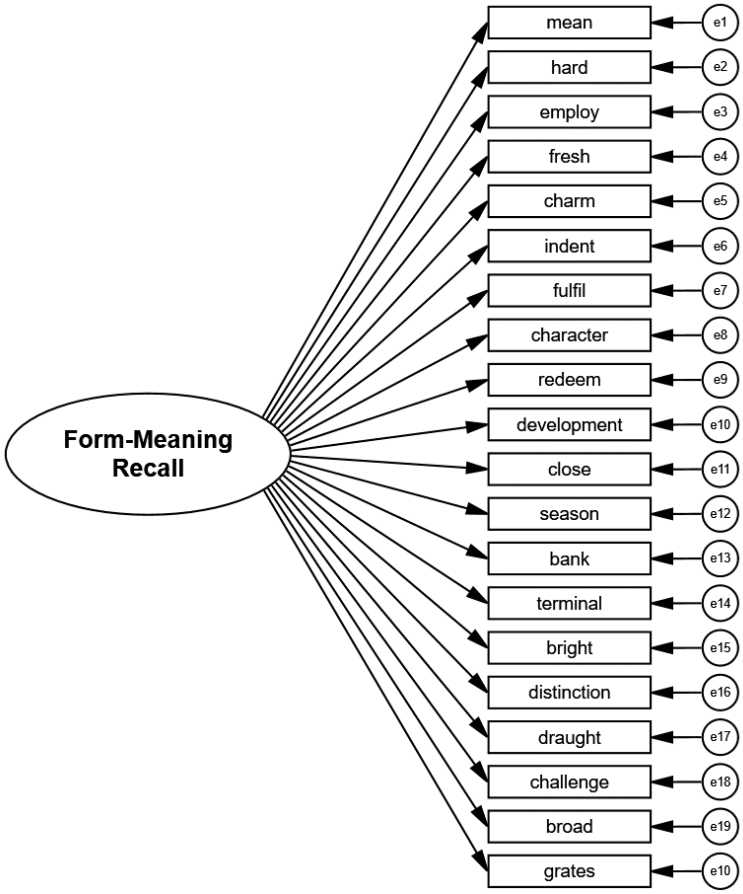
6.5.2.1. Validation of measurement instruments with individual words

It is best practice in SEM analysis to run Confirmatory Factor Analysis (CFA) for the measurement models in order to validate the latent factors before modelling the relationships between them. Therefore, I conducted a series of CFA to validate the word knowledge aspects constructs prior to testing the hypothesised structural model.

Each recognition and recall word knowledge aspect was measured by 20 target words. However, due to the relatively small sample size and the parsimony of the model, these recognition and recall aspects were conceptualised in the hypothesised model as the result of the total test score for the 20 items, instead of the individual words' scores. Nevertheless, in order to ensure that the tests used to measure these aspects were valid and consistent with my understanding of the construct, I first ran independent CFA for

each aspect including the individual words. In each of these CFA, the results of the 20 target words used to measure each of the recognition and recall aspects were loaded as indicators onto their respective word knowledge component, which were described as latent variables (see Figure 15 for an illustration). Eight CFA models were tested, one for each of the word knowledge aspects. Because the scores for each word were categorical, the estimator WLSMV was used (robust version of WLS estimator, which is more appropriate for categorical, non-normally distributed variables).

Figure 15 Example of CFA for each word knowledge component.



For the sake of space, only the model fit indices and reliability of the CFA models are reported. Table 45 shows that the different measures of word knowledge aspects achieved good reliability and model fit coefficients, suggesting that the target words are measuring the construct they were designed to measure and confirming that the different vocabulary components are valid psychometric constructs.

Table 45 Fit indices and reliability for the CFA models with 20 target words as indicators (Spanish group).

	Model fit indices						
	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	Reliability
Acceptable fit			>.05	1-3	>.95	<.05 / <.08	α >.7
Form–Meaning Recall model	189.717	170	.143	1.12	.97	.03 (.000-.049) *p* = .962	α = .91
Form–Meaning Recognition model	169.885	135	.023	1.26	.95	.04 (.017-.061) *p* = .725	α = .91
Derivative Recall model	219.316	170	.006	1.30	.99	.05 (.025-.061) *p* = .673	α = .96
Derivative Recognition model	229.898	170	.002	1.30	.99	.05 (.032-.065) *p* = .501	α = .95
Multiple-Meanings Recall model	224.271	170	.003	1.32	.98	.05 (.028-.063) *p* = .6	α = .94
Multiple-Meanings Recognition model	228.112	170	.002	1.34	.96	.05 (.031-.065) *p* = .53	α = .91
Collocation Recall model	214.374	170	.012	1.26	.95	.04 (.021-.059) *p* = .75	α = .90
Collocation Recognition model	196.776	170	.078	1.16	.95	.03 (.000-.052) *p* = .93	α = .90

Some features of these models need to be explained further. First, for some models (i.e., Form–meaning Recognition, Derivative Recall, Derivative Recognition, Multiple-Meanings Recall, Multiple-Meanings Recognition and Collocation Recall) the chi-square *p* value appears as significant. This does not indicate, however, that the models are not suitable, since it is possible to have a valid model even if the chi-square is significant (Kenny et al., 2014).

The reason is that chi-square is overly sensitive to sample size and non-multivariate normality of the data (Bentler & Bonett, 1980), being able to inflate its values and making this index less appropriate for model evaluation in these conditions. Strict interpretations of the chi-square and its significance level may underestimate the goodness-of-fit of a model (Byrne, 2016). As such, many researchers claim that chi-square should not be considered the basis for model acceptance or rejection (e.g. Awang, 2012; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Vandenberg, 2006).

That is, with good fit indicated by other fit indices (such as CFI or RMSEA), a significant chi-square is not a good-enough reason to discard or modify a model, leading some researchers to recommend ignoring the chi-square *p* value when reporting measurement model fit (Awang, 2012). Therefore, given the non-multivariate normality of the dataset and the relatively small sample size in this study, and based on the other model fit indices, I consider these models valid even with the significant chi-square.

Second, the lower number of degrees of freedom for the Form–meaning Recognition model responds to the correlation of some measurement errors. The evaluation of this model suggested that error correlation or item deletion were desirable to achieve a better model fit. Removing worse-performing target words would cause a loss of valuable information for the purpose of this study, and thus, error correlations, although not perfect, were considered the best solution. The model fit indices and reliability suggested that the model was suitable with such modifications.

Finally, the internal structure of the Form–meaning Recall and Collocation Recognition models showed that some target words did not achieve significance or had a lower-than-recommended factor loading. For example, the words *close*, *bank* and *character* were not-significant and had low factor loadings in the Form–meaning Recall model (see Appendix 9). Although it is generally recommended that any item with factor loading of less than .40 should be deleted (Hooper et al., 2008), I did not consider it appropriate in this study for a number of reasons. First, the target words were selected so that some were easier than others in the different components (i.e., different frequencies). Thus, the participants' performance was expected to be different for easier and more difficult words, with most learners knowing the easiest and only few learners knowing the most difficult ones. Moreover, different words have different difficulties in the various components (e.g., the multiple meanings of the target word *bank* are expected to be more difficult than that of the word *hard*). Therefore, some words were expected to be generally known better than others in a specific aspect. As a consequence, due to the purpose of the study and the nature of the data, removing non-significant or low-factor-loading items would affect the integrity of the measures. Finally, low-factor loadings simply indicate a useless item (Awang, 2012), and are recommended to be removed because they can affect the model fit. However, in these models both the reliability and the model fit were good, and thus, I decided not to exclude them from the analysis.

In order to test the hypothesised model, the results for the individual target words needed to be comprised into a composite total score, and this total score used as indicator of the word knowledge components. This reduces any potential impact of the above issues for the general model.

To sum up, despite the issues mentioned above, Table 45 illustrates that the measurement instrument employed for each word knowledge aspect are reliably assessing the construct they were designed to assess, and thus are appropriate representations of the word knowledge components.

6.5.2.2. Validation of latent variables for hypothesised model

Once the validity of the different aspects was checked individually, I validated the latent variables as represented in the hypothesised model. In order for the hypothesised model to be identified the recognition and recall aspects of each word knowledge component were specified as the total test score for the 20 target words used to measure them[17]. This way, the recognition and recall composite scores were conceptualised as individual indicators of the word knowledge components. Thus, the model was comprised of the four word knowledge components as inter-related latent dimensions of vocabulary knowledge, each tapped by the recognition and recall indicators (see Figure 16). Again, in order to examine the belonging of the composite recognition and recall indicators to their respective factors, I first evaluated the fit of the measurement models by conducting a pooled-CFA of all the factors included in the hypothesised model[18]. Because the variables in the models were continuous and non-normally distributed (as shown by Mardia's Multivariate Normality Test (Mardia, 1970)), I used the MLR estimator. This pooled-CFA was used to assess the unidimensionality, validity and reliability of each latent construct before including it in the hypothesised SEM model.

[17] The inclusion of the 20 target words as indicators of each word knowledge component was impossible because the number of data points was less than the number of parameters to be estimated in the model, and thus the model could not be identified (Byrne, 2016). A much larger sample size would be required for such a model to achieve identifiability. As a solution, the composite total score for the 20 target words in each word knowledge aspect was employed as continuous indicators.

[18] Running individual CFA's for each measurement model (i.e., a factor and its indicators) was not possible due to the limited number of indicators (two). A pooled-CFA including all the factors in the hypothesised model was run instead.

Figure 16 Pooled-CFA of all factors in hypothesised model (Spanish group).

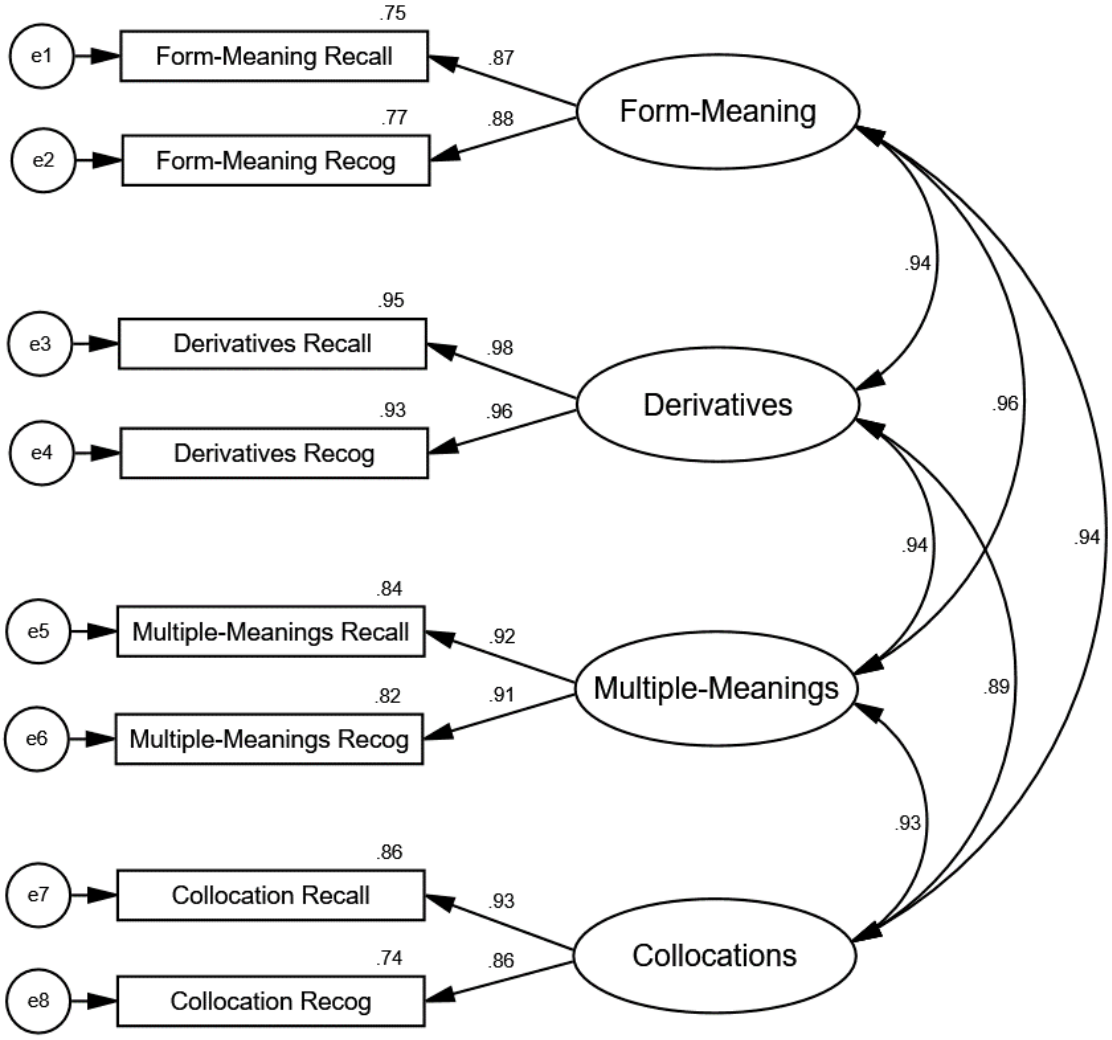


Table 46 shows the fit indices of the Pooled-CFA. The model's χ^2 was 24.368 (*df* = 14; *p* = .04), and the ratio of χ^2/df was 1.74, which indicates good fit. The RMSEA of .07 was better than the standard for acceptable fit (< .08), the CFI of .99 indicate almost perfect fit (i.e., near 1.0). Finally, the SRMR of .012 was very close to 0, and thus shows good fit.

Table 46 Pooled-CFA fit indices (Spanish group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	24.368	14	.041	1.74	.99	.07 (.017-.117) *p* = .201	.012

The fit indices reached the threshold to consider the model valid, and thus construct validity was accomplished. There is only one value that is not as high as might be hoped: the chi-square *p* value, which is just significant. Nevertheless, research has found that it is the most sensitive index to my sample size and type of data, and thus the model is still valid based on the rest of indices. Unidimensionality was achieved by each factor (factor loadings >.40). Convergent validity was also attained by every factor, as represented by Average Variance Extracted (AVE) (Form–meaning AVE = .77, Derivatives AVE = .94, Multiple-Meanings AVE = .84, Collocation AVE = .80). Finally, the overall model was found to be reliable (Cronbach's α = .90), as well as the measurement models individually (Form–meaning Cronbach's α = .84, Composite Reliability (CR) = .87, Derivatives Cronbach's α = 0.97, CR = .97, Multiple-Meanings Cronbach's α = .91, CR = .91, Collocations Cronbach's α = .89, CR = .89) (see Table 47).

Table 47 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for Pooled-CFA model (Spanish group).

Factors	**AVE**	**CR**	**α**
Acceptable value	>.5	>.6	>.7
Form–Meaning link	.77	.87	.84
Derivatives	.94	.97	.97
Multiple-Meanings	.84	.91	.91
Collocation	.80	.89	.89
Overall model			.90

Table 48 shows the factor loadings, squared multiple correlations, *z* scores and significance for the pooled-CFA model. Since all paths were significant, convergent validity was confirmed. The squared multiple correlation indicates that the reliability of the indicators with composite scores is good (>.70), and by extension, the factors are reliable constructs. CFA estimates correlations between the learners' scores in each test on the vocabulary knowledge latent dimension. These correlations in a CFA are a better representation of relationships than the correlations conducted between individual measures (see Table 44). The reason for this is that the correlations in CFA are not negatively influenced by measurement error associated to the tasks. As a consequence,

they are better estimates of the relationships between the components. These true scores' correlations between the four dimensions were all statistically significant, and very high (see Table 48).

Table 48 *Z* scores and significance for Pooled-CFA (Spanish group).

Paths	**Factor Loadings R**	**R^2**	***Z***
Form–Meaning → Form–Meaning Recog	.88	.77	14.267***
Form–Meaning → Form–Meaning Recall	.87	.75	15.291***
Derivatives → Derivatives Recog	.96	.93	22.394***
Derivatives → Derivatives Recall	.98	.95	21.395***
Multiple-Meanings → MM Recog	.91	.82	16.936***
Multiple-Meanings → MM Recall	.92	.83	14.863***
Collocation → Collocation Recog	.86	.74	13.784***
Collocation → Collocation Recall	.93	.86	17.766***
	Correlations (R)		***Z***
Form–Meaning ←→ Derivatives	.94		39.429***
Form–Meaning ←→ Multiple-Meanings	.96		37.755***
Form–Meaning ←→ Collocations	.94		36.208***
Derivatives ←→ Multiple-Meanings	.94		52.969***
Derivatives ←→ Collocations	.89		41.635***
Multiple-Meanings ←→ Collocations	.93		41.465***

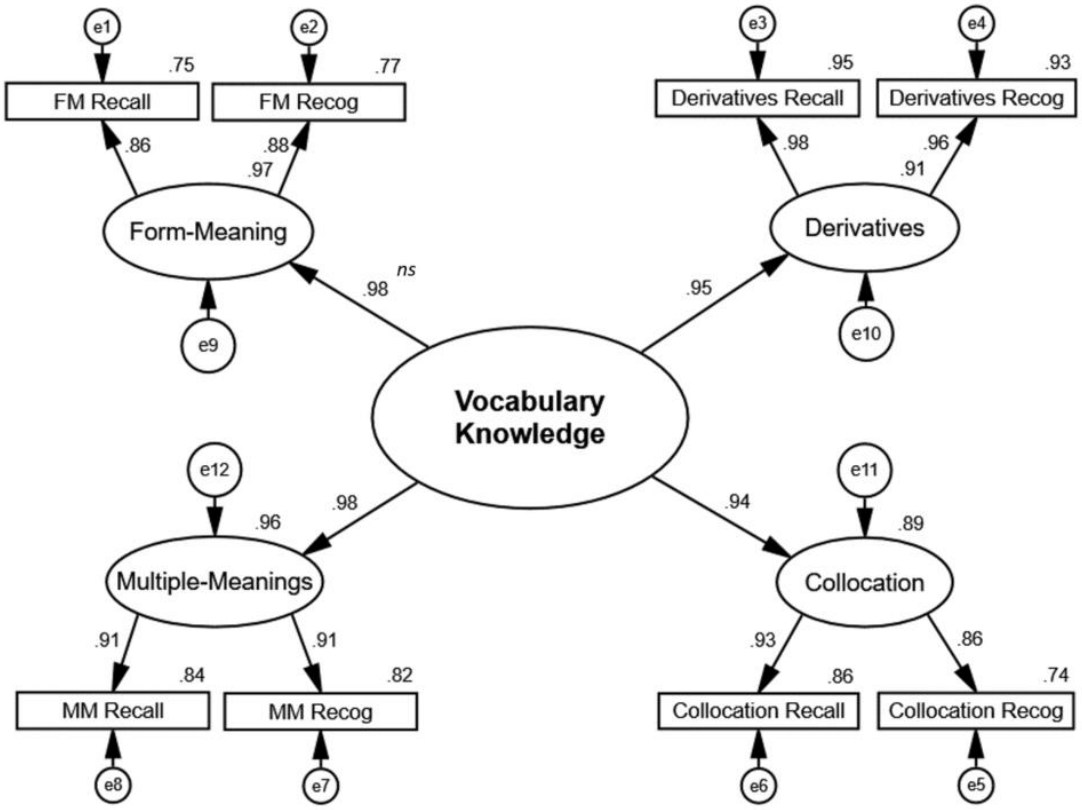
*** p <.001.

The Pooled-CFA results show that the indicators reliably represent the concept of each word knowledge components, and thus these components are suitable to be built into the hypothesised model. Nevertheless, the extremely high intercorrelations among the different factors (.89-.96) suggest that they are representing very similar concepts. This may indicate redundancy, and thus, lack of discriminant validity between the factors (Kline, 2016). This will be discussed in the hypothesised model evaluation.

6.5.2.3. Hypothesised model

The hypothesised model was fitted to SEM employing the MLR estimator. This model symbolises a second-order structural model in which the overall vocabulary knowledge construct was explained by the four word knowledge factors, which were comprised of their combined recognition and recall masteries. Two indicators per latent variable are enough in this model because the factors were shown to correlate with each other (McDonald & Ho, 2002). Figure 17 shows the strength of the relationships between word knowledge components and vocabulary knowledge as described in such model.

Figure 17 Model 1: Hierarchical model of Vocabulary Knowledge (Spanish group).



Note: All values $p<.01$, except *ns*=nonsignificant ($p=.14$).

Table 49 presents the fit indices for the hypothesised model. The χ^2 for the model was 26.426 ($df = 16$; $p = .05$), and the ratio of χ^2/df was 1.53, which indicates good fit. RMSEA of .06 was near the standard for close fit (<.05) and better than the standard

for acceptable fit (<.08), the CFI of .99 not only met the standards for acceptable fit (>.95), but also indicate almost perfect fit (i.e., near 1.0). Finally, SRMR of .013 was close to 0, showing very good fit. Since all the model's fitness indices met the required acceptability level, construct validity was achieved.

Table 49 Model 1: Fit indices (Spanish group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	26.426	16	.05	1.53	.99	.06 (.010-.111)	.013
						p = .24	

All factor loadings were way above .40, indicating that unidimensionality was achieved. Convergent validity was also attained, as represented by AVE of the overall Vocabulary Knowledge construct (AVE = .93) and the four sub-constructs (Form–meaning AVE = .77, Derivatives AVE = .94, Multiple-Meanings AVE = .84, Collocation AVE = .80). The overall model was found to be reliable (Cronbach's α = .90, CR = .98), as well as the measurement models individually (Form–meaning Cronbach's α = .84, CR = .87, Derivatives Cronbach's α = .97, CR = .97, Multiple-Meanings Cronbach's α = .91, CR = .91, Collocations Cronbach's α = .89, CR = .89).

Table 50 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for hypothesised model (Spanish group).

Factors	AVE	CR	α
Acceptable value	>.5	>.6	>.7
Form–Meaning link	.77	.87	.84
Derivatives	.94	.97	.97
Multiple-Meanings	.84	.91	.91
Collocation	.80	.89	.89
Overall model	.93	.98	.90

Table 51 Relative contribution of the different aspects to vocabulary knowledge, *Z* scores and significance (Spanish group).

Paths	**Factor Loadings R**	**R^2**	***Z***
Form–Meaning → Form–Meaning Recall	.86	.75	1.510 (*p* =.131)
Form–Meaning → Form–Meaning Recog	.88	.77	1.497 (*p* =.134)
Derivatives → Derivatives Recall	.98	.95	9.009***
Derivatives → Derivatives Recog	.96	.93	9.019***
Multiple-Meanings → MM Recall	.91	.84	2.676**
Multiple-Meanings → MM Recog	.91	.82	2.668**
Collocation → Collocation Recall	.93	.86	6.412***
Collocation → Collocation Recog	.86	.74	6.439***
	Standardised Regression weights		***Z***
Vocabulary Knowledge → Form–Meaning	.98		1.466 (*p* =.143)
Vocabulary Knowledge → Derivatives	.95		8.076***
Vocabulary Knowledge → Multiple Meanings	.98		2.554**
Vocabulary Knowledge → Collocations	.94		5.846***

** $p < .01$, *** $p < .001$.

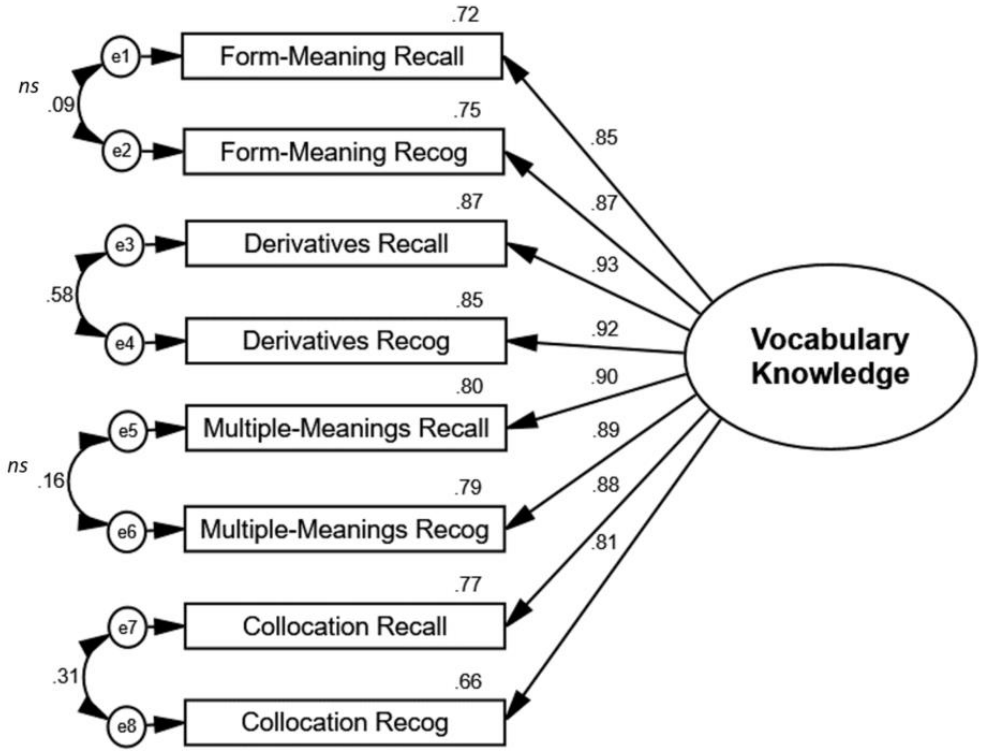
Thus, Model 1 shows a good overall fit with the data. However, an examination of the strength of the relationships in the model showed that the regression coefficients (i.e., paths between Vocabulary Knowledge and the four word knowledge components) were all very high (β = .94-.98), suggesting lack of discriminant validity and affecting the significance of some paths (see Table 51). These β coefficients indicate that all these components are highly intercorrelated, and thus, empirically indistinguishable, so they can be better understood as a single construct. That is, the regressions between them were too high to claim that they were different dimensions of vocabulary knowledge. This is in line with the correlations in Table 48 which show that all of the aspects were strongly interrelated and the proportion of variance in each variable explained by vocabulary knowledge (i.e., number on top of each oval), which ranges between 89%

and 97%. Therefore, even though the fit indices for Model 1 were acceptable, the high regressions between the components and Vocabulary Knowledge suggest that they are redundant and comprise a single construct, and therefore interpretation of the coefficients is meaningless. In other words, this model is not well supported by the data, and cannot be considered a good representation of vocabulary. Respecification of the hypothesised model is required removing the indiscriminant variables (Hooper et al., 2008).

6.5.2.4. Revised model

Section 6.2.3 describing the steps in SEM analysis shows that the model evaluation stage can provide information about whether a hypothesised model needs to be respecified. Since Model 1, which divided vocabulary knowledge into multiple dimensions, was not supported by the data due to lack of discriminant validity, some modifications of this hypothesised model were needed. I reconceptualised the model based on the results from the IS in Chapter 4, which showed that recognition and recall knowledge in general behaved quite differently from each other. Therefore, it was remodelled considering the recognition and recall aspects as individual direct indicators of the general vocabulary knowledge construct. Thus, the redundant latent factors for each component were eliminated. Moreover, following previous findings (e.g., Laufer & Paribakht, 1998) I anticipated that recognition and recall aspects of the same component (e.g., Derivative Recall and Derivative Recognition aspects) would correlate. Since CFA has been found to be a good method for investigating the dimensionality of a construct (Kieffer & Lesaux, 2012a; Scott et al., 2017), I employed CFA to test the fit of the data to the revised one-dimensional model. This model is illustrated and analysed in Figure 18.

Figure 18 Model 2: Recognition and recall knowledge as independent aspects of vocabulary knowledge (Spanish group).



Note: All values $p<.001$, except *ns*=nonsignificant ($p=.42$ and .09).

Table 52 Model 2: Fit indices (Spanish group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	22.078	16	.14	1.16	.99	.05 (.000-.099) p = .44	.013

The model's χ^2 (22.078; df = 16; p = .14) and the ratio of χ^2/df (1.16) show a good fit. RMSEA of .05 shows close fit (< .05), as well as the CFI of .99 that indicate almost perfect fit. SRMR of .013 was very close to 0. All the model fit indices reached and exceeded the commonly accepted thresholds, supporting the suitability of the revised model, and thus, the validity of the construct.

When comparing the fit of the hypothesised and revised models, the parsimony fit indices (AIC and BIC) do not seem to be different, suggesting both models are

equally good in terms of parameter estimation. However, the revised model shows an improvement from the hypothesised model, particularly in the χ^2, its *p* value, χ^2/df and RMSEA. This indicates that Model 2 has an overall superior fit, and thus appears to be a better representation of vocabulary knowledge, based on my data.

Table 53 Model comparison for parsimony (Spanish group).

Model	AIC	BIC	Sample-sized adjusted BIC
Model 1	6,247.37	6,330.53	6,241.93
Model 2	6,247.37	6,330.53	6,241.93

As in the previous model, the assumptions of unidimensionality, convergent validity and reliability were tested. All factor-loadings are positive and above the appropriate level (>.40), which suggests the model achieved unidimensionality. Besides, all loadings are statistically significant and AVE exceeds .50 (AVE = .78), showing that convergent validity is achieved by this model (Awang, 2012). The reliability and internal consistency of the model was checked by means of Cronbach's α and Composite Reliability (CR). For Model 2, both Cronbach's α and CR showed to be very good (α =.97, CR = .97 (CR >.60 is considered good, (DeVellis, 2012)). This suggests that the vocabulary knowledge construct as explained by those aspects is a valid and reliable construct. In sum, the relationships between components were satisfactorily verified and significant, demonstrating the validity and reliability of the model.

Table 54 shows the standardised factor loading and the square multiple correlations to see the variance explained by each aspect.

Table 54 *Z* scores and significance of word knowledge aspects as indicators of vocabulary knowledge (Model 2) (Spanish group).

Paths	**Factor Loading R**	**R^2**	***Z***
Vocabulary Knowledge → FM Recall	.85	.72	14.499***
Vocabulary Knowledge → FM Recog	.87	.75	14.048***
Vocabulary Knowledge → Deriv Recall	.93	.87	19.753***
Vocabulary Knowledge → Deriv Recog	.92	.85	20.654***
Vocabulary Knowledge → MM Recall	.90	.80	14.155***
Vocabulary Knowledge → MM Recog	.89	.79	16.042***
Vocabulary Knowledge → Collo Recall	.88	.77	16.377***
Vocabulary Knowledge → Collo Recog	.81	.66	12.833***
	Error Correlations		***Z***
FM Recall ←→ FM Recog	.09		0.80 *p*=.42
Derivatives Recall ←→ Derivatives Recog	.58		4.564***
MM Recall ←→ MM Recog	.16		1.275 *p*=.09
Collocation Recall ←→ Collocation Recog	.31		3.425***

*** $p < .001$.

The *R* values show that all the word knowledge aspects load highly (>.40, Hooper et al., 2008) on the vocabulary knowledge latent construct. These *R* values range from .81 (R^2 = .66) to .93 (R^2 = .87). Thus, the variance explained ranges from 66% to 87%. The lowest factor loading is found for collocation recognition knowledge (*R* = .81), which is still a very strong correlation. This suggests that when we take all these vocabulary knowledge aspects into account, recognition collocation knowledge explains 66% of overall vocabulary knowledge. The highest correlations are between vocabulary knowledge and the derivative aspects (Derivative Recall *R* = .93, Derivative Recognition *R* = .92), which suggest that the greatest amount of variance in vocabulary knowledge is explained by knowledge of derivatives (R^2 = .87, R^2 = .85). In a study with native speakers of English, Tannenbaum et al. (2006) found that recall knowledge of polysemy had the strongest contribution to depth of vocabulary (.69), when compared to other aspects such as recall knowledge of associations (what they called semantic category). In this model, recall knowledge of multiple meanings was also one of the strongest contributors to word knowledge for L2 learners (*R*=.90). An assumption of CFA is that if the indicators belong to a unique construct, as in this case, then these

indicators are very strongly correlated with each other (Awang, 2012). Therefore, overall the high factor loadings show that the independent word knowledge aspects intercorrelate very strongly through the vocabulary knowledge construct, confirming their belonging to an underlying construct.

In general, the recall aspects tend to contribute to vocabulary knowledge slightly more than their recognition counterparts, with the exception of the Form–meaning Recall/Recognition pair. Nevertheless, the factor loadings are high for all word knowledge aspects, suggesting that all of them contribute in a similar manner to, and are explained by, a unique factor: Vocabulary Knowledge.

Finally, regarding the correlated errors, only the measurement error correlation for the derivatives aspects and the collocation aspects were significant (r = .58, p = .000, and r = .31, p = .003, respectively, see Table 54). This suggests that these two components are measuring something is common other than the latent construct (Kline, 2016). In the case of derivatives, the correlation suggests that they are responsible for 33% of some unexplained variance in the model. This might be a reflection of the grammatical and morphological information comprised in knowing derivatives. In order to understand, and even more to produce, derivations of a word, learners need to have metalinguistic knowledge of the word formation rules in both the L1 and the L2, as well as knowledge of the specific affixes to be applied to particular word classes (Scott, Miller, & Flinspach, 2012). In the case of collocations, the unexplained variance they measure is just below 10%, and may refer to the contextual knowledge required to use a word appropriately. Therefore, these significant correlations may be an indication of the different types of linguistic knowledge needed to use different aspects of vocabulary appropriately in context. In turn, the lack of significant error correlation between the other two pairs of aspects (Form–meaning and Multiple-Meanings) suggests that the unique variance between the recognition and recall masteries is explained only by knowledge of the overall vocabulary knowledge construct. This may mean that these two components are primarily meaning-based, and this seems to be the essence of lexical knowledge.

In sum, Model 2 shows to be a good representation of vocabulary knowledge. All word knowledge aspects make a large contribution to the explanation of the Vocabulary Knowledge construct, which demonstrates they are all essential components of knowing vocabulary.

6.5.2.5. Alternative models

The nature of the present study is exploratory in that it is one of the first attempts to empirically investigate the overall vocabulary knowledge construct. Since there is not generally accepted theory of relationships between word knowledge aspects, I wanted to test other theoretically valid alternatives. Thus, along the hypothesised model built to test the multidimensionality of vocabulary, I was interested in checking other models based on previous findings or claims.

First, the hypothesised model was designed to check that the word knowledge components belonged to the vocabulary construct, and thus, confirm the multidimensionality of vocabulary. However, it could not inform about possible predictive relationships between aspects. Thus, a model examining the predictive effect between the form–meaning link component and the rest of the aspects was conceptualised. This model was based on the idea that the number of words a learner knows (i.e., size of vocabulary knowledge) can be an indicative of the quality of word knowledge (i.e., depth) (Schmitt, 2014). Figure 19 shows the conceptualisation of such model.

Figure 19 Direct effect of form–meaning link on depth of knowledge aspects (Spanish group).

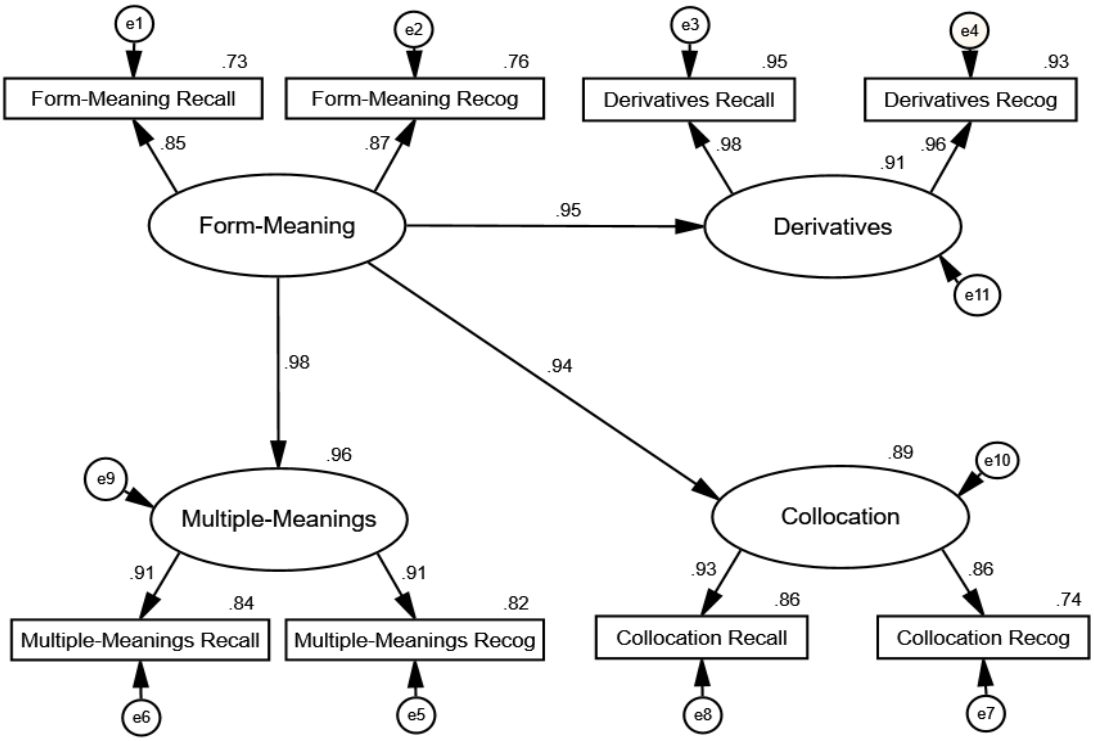


Table 55 Model fit for Alternative Model 1 (Figure 19) (Spanish group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	24.368	14	.04	1.74	.99	.07 (.017-.117) *p* = .20	.012

Table 55 shows a good overall model fit in terms of fit indices for the alternative model presented in Figure 19. However, I found the same problem as in the hypothesised model. The high regression coefficients (β = .94-.98) indicate that there is no distinction between the different components as independent dimensions. Therefore, although the model is theoretically possible, it was not empirically supported by my data.

Second, I was interested in whether the amount of overall recognition vocabulary was indicative of the recall knowledge of a learner. Thus, a model was conceptualised in which the latent construct of recognition vocabulary knowledge, comprised of the recognition tests of each component, predicted recall vocabulary knowledge, formed by the four recall tests (Figure 20).

Figure 20 Direct effect of recognition on recall masteries (Spanish group).

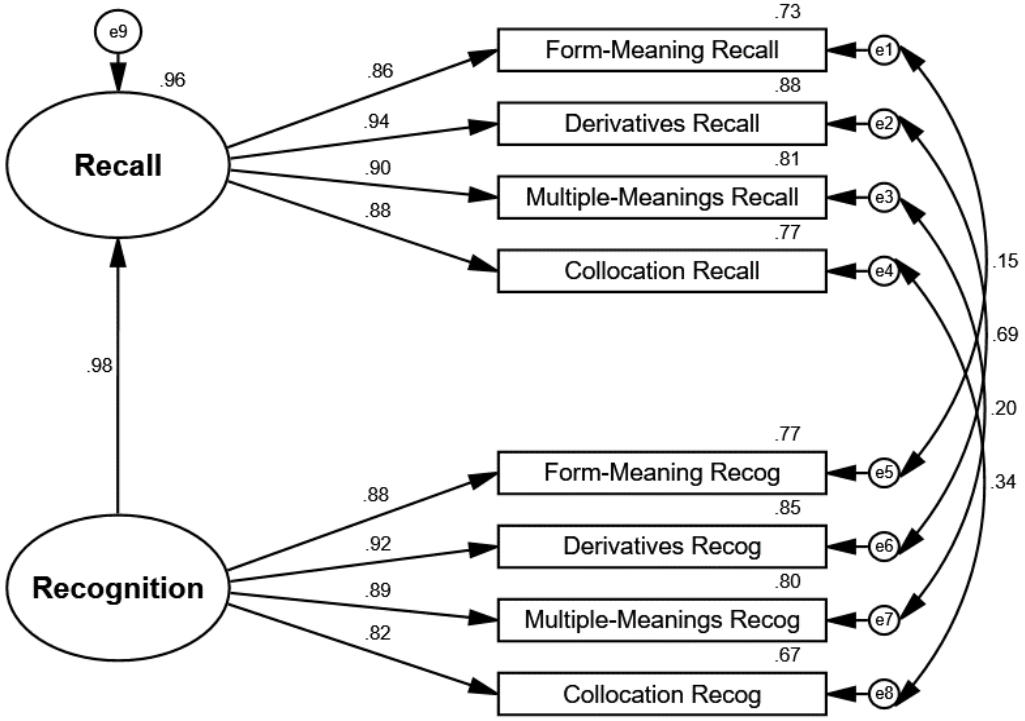


Table 56 Model fit for Alternative Model 2 (Figure 20) (Spanish group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	12.890	15	.61	.86	1	.00 (.000-.068) *p* = .86	.011

Again, Table 56 shows a good fit for Alternative Model 2, but the regression coefficient between the two latent variables revealed absence of discriminant validity between them (β =.98). Thus, if these components cannot be discerned from one another, they lose all their predictive power.

Although these models illustrate interesting theoretical relationships about vocabulary knowledge, the finding of lack of discriminant validity between the word knowledge aspects shown by the hypothesised model implies that these models cannot have an explanatory purpose. That is, if the word knowledge components cannot be considered independent factors, but aspect of a common construct, predictive relationships between them cannot be described. Therefore, these findings imply that the aspects are so tightly interconnected that we cannot reliably discriminate them, and hence the predictive effect of one over another is meaningless (i.e. "*X* predicts *X*"). The findings from these alternative models are further evidence for the lack of multidimensionality of vocabulary knowledge, and thus give strength to the unidimensional model of vocabulary knowledge.

6.6. Discussion

The current study is the first attempt to empirically explore the overall nature of L2 vocabulary knowledge following Nation's (2013) framework of word knowledge components. This framework has been very influential, and has helped researchers and practitioners conceptualise vocabulary knowledge for a generation. My study enhances this framework by giving an indication of how the components relate to one another and to the whole vocabulary knowledge construct. However, my findings also challenge

some typical views on vocabulary knowledge, in particular related to its multidimensional nature[19]. Pedagogical implications will be discussed in Chapter 8.

6.6.1. Representation of word knowledge aspects

The different aspects of word knowledge can be operationalised in multiple ways. There is no agreement to which is the best test format to represent each specific concept, and, as a result, no standardised test exists to measure the particular types of knowledge, except arguably for the form–meaning link (e.g., VST or VLT). Therefore, it is fundamental in a study attempting to empirically describe a theoretical construct that the instruments designed to assess each aspect of word knowledge are valid and reliable. The results from the individual CFA for each of the eight word knowledge tests confirm that the test battery employed in this study is reliably measuring the specific vocabulary components, suggesting that the formats and items chosen for this purpose were appropriate. That is, the findings provide validity for the different tests, and thus, the aspects they represent.

If the overall tests are valid, the items that comprise them are so by extension. Nevertheless, there are always better and worse performing items in a test. In this battery, some words in two tests (namely form–meaning recall and collocation recognition) showed very limited effect in explaining the particular concept under study. This situation reflects a reality in L2 vocabulary learning: some words are easier to learn than others, in general and in specific components (Webb & Nation, 2017). Frequency in language is a characteristic of words usually employed in vocabulary study to predict the probability of knowing a word (Schmitt & Schmitt, 2014). More frequent and/or easier words are expected to be known by most people, and thus, less variation in the responses is to be expected, which in turn affects the strength of relationship of those items. Although the participants in this study had various L2 proficiency levels in order to ensure variability in responses, some easier words were inevitably known by most participants (e.g., *close* in the form–meaning link component) or not known by almost any participant (e.g., *redeem* or *indent* in the collocation

[19] The interpretation of the findings from this study does not indicate that Nation's description of vocabulary is wrong. Rather, it suggests that the way the vocabulary knowledge aspects are learnt and interrelate is dissimilar from the descriptions of vocabulary components as independent constructs. Thus, this discussion refers to vocabulary *knowledge*, not the linguistic area of vocabulary and its intrinsic characteristics.

recognition aspect). This anticipated general easiness/difficulty of these words regardless of the participants' proficiency level made them worse statistical representations of the type of knowledge tested. Nevertheless, their inclusion in the test and analysis is essential in order to obtain a more representative description of knowledge of a vocabulary component. Since the model fit was good for each word knowledge component regardless of these low strength items, the findings give validity to my measures as comprehensive indicators of the word knowledge components.

The reliability of the individual recognition and recall measures was confirmed when the discrete scores for the 20 target words were summed in a single indicator. The pooled-CFA also showed that the correlations between the word knowledge components as latent dimensions, comprised of their combined recognition and recall scores, were very high, all ranging between .89 and .96. These high correlations question the independence of the four vocabulary components (form–meaning, derivatives, multiple meanings and collocations), suggesting that they should not be treated as completely independent constructs (Kenny, 2016). Rather, the various aspects are related and provide complementary information about the learner's depth of knowledge, and thus none can replace each other.

These findings have some applications to the assessment of word knowledge aspects. Given that construct validity has been demonstrated for the test formats employed in this study as indicative of knowledge of vocabulary components, they could be used as the basis for the development of standardised tests of word knowledge components and overall vocabulary knowledge. The creation of standardised tests would enable better comparability of research on depth of knowledge, as well as more comprehensive understanding of this construct (Fitzpatrick & Milton, 2014).

6.6.2. Unsupported multidimensionality of vocabulary knowledge

Nation's (2013) framework suggests that both receptive and productive knowledge are part of each vocabulary component, which in turn make up the greater vocabulary knowledge construct. Thus, in this conceptualisation, vocabulary is understood as a higher-order construct comprised of multiple sub-dimensions, where mastering one of those dimensions (e.g., form–meaning link) involves acquiring its receptive and productive knowledge. However, when this conceptualisation was

modelled, I found that the four latent word knowledge components loaded on the vocabulary knowledge construct extremely strongly and almost identically (β =.94-.98). This finding relates to the implicational scale described in Study 1 (Chapter 4), where all of the recognition knowledge aspects were found to behave similarly to each other. The same was true for the recall aspects. This meant that no word knowledge component was evidenced both in recognition and recall before another component. Rather, all of the recognition knowledge aspects were mastered before the recall ones. Thus, combining the recognition and recall tests as indicators led to the four vocabulary components being known to a similar degree, and therefore they could not be discriminated using the model. This result indicates that the latent components cannot be considered independent concepts, and thus the so-called different dimensions of vocabulary seem to be artificial boundaries in the mental lexicon. This finding was also discovered in the alternative vocabulary models, where the components could not be related in terms of predictability because they did not show discriminant validity. As a consequence, my study provides empirical evidence to claim that the multidimensional understanding of vocabulary knowledge does not seem to be an accurate representation of this construct in practice. In other words, the vocabulary knowledge construct is not multidimensional in nature.

These surprising results oppose not only the most widely accepted theoretical conceptualisations of vocabulary, but also the two preceding studies that claimed the multidimensionality of vocabulary knowledge employing SEM (Kieffer & Lesaux, 2012a; Scott et al., 2017). However, there are certain differences between those studies and the present one that can explain the opposing results. The main distinction is the conceptualisation of vocabulary followed by each study. While the selection of vocabulary components in my study is based on the most widely accepted understanding of word knowledge in the L2 vocabulary field, the previous studies do not seem to follow a specific framework. Rather, they combine aspects that various independent studies have considered to be important features related to vocabulary and reading (e.g., Carlo et al., 2004; Kieffer & Lesaux, 2008; Nagy et al., 2006). Conversely, my selection of word knowledge components was ruled by having a representative of the three main features of vocabulary: form, meaning and use. Although this is not in essence a disadvantage, I believe it makes their selection of vocabulary components less principled and comprehensive. For example, none of these studies include collocational

knowledge as an aspect of vocabulary. Similarly, they only measure recognition of the different aspects, and in no case recall.

Another difference is with the specific measures employed. Kieffer and Lesaux (2012a) used both standardised tests and research-designed tasks developed to study aspects of lexical knowledge that are reading related, since their focus is on the application of vocabulary for reading. Thus, apart from assessing different sets of words (e.g., in one test only academic words, in another test non-words), these tests were not originally designed to be employed concurrently, and thus their operationalisation already targeted independent dimensions. This provided more diverse measures that are likely to behave in dissimilar ways, be less comparable, and thus result in empirically distinct constructs. As for Scott et al.'s (2017) study, although their measurement instruments were conceptualised to be tested together, their decision of only including in the analysis four of the original 24 target words limits their conclusion of multidimensionality. It can be that the selected words behave the most differently across the various components, and thus the multidimensionality only holds for these items. It would be interesting to see what happens when adding the information about all the target words in the analysis, as in the case of my study. It may be that the results confirm mine in this situation. Finally, and more importantly, Scott et al. (2017) did not provide a direct check of convergent or discriminant validity. My multidimensional model, as theirs, provided a good fit, but was discarded due to the lack of empirical discrimination between dimensions. Thus, their multidimensional claim is dependent on the availability of such information.

Overall, these differences between studies led to the conflicting findings. However, the limitations of the previous research suggest that its results are not conclusive, strengthening my claim that multidimensionality of vocabulary knowledge is not empirically supported by present data. I believe the specific characteristics of my study and data make the claim that vocabulary knowledge is unidimensional theoretically and empirically sound, and therefore more appropriate for the conceptualisation of vocabulary.

6.6.3. Unidimensional conceptualisation of vocabulary knowledge

The revised model conceptualised vocabulary knowledge as a unique dimension comprised of the independent recognition and recall masteries of the word knowledge components tested. This model not only met the evaluation criteria regarding construct validity, with an extremely good fit, but also in terms of convergent and discriminant validity, unlike the multidimensional model. Therefore, it is empirically a better representation of vocabulary knowledge, suggesting this is the way it should be conceptualised.[20] Although at first this description of vocabulary may seem a little counterintuitive, it is actually consistent with the general results of most previous research which show very high correlations between components (e.g., Chui, 2006; Gyllstad, 2009; Milton & Riordan, 2006; Vermeer, 2001). My study provides empirical evidence, with more appropriate statistical techniques, to make those claims. It also confirms the findings by Spencer et al. (2015) that different types of vocabulary knowledge interrelate so highly that they comprise a unidimensional construct, where all of the various kinds of knowledge are essential facets of vocabulary. In fact, previous studies that showed different vocabulary dimensions (e.g., Kremmel et al., 2017; Tannenbaum et al., 2006) anticipated the unidimensionality of vocabulary knowledge when applying the discriminant validity assumption (i.e., correlations between lexical aspects of .90 and .88 respectively).

This unidimensional model still suggests that vocabulary has a multifaceted nature in the sense that its various interrelated knowledge types are different in the difficulty they pose to L2 learners. However, my model shows that these types of knowledge do not constitute independent dimensions, since this term brings implications of them being fundamentally different from each other. Rather, vocabulary knowledge should be understood as a unique dimension that involves various highly-interconnected aspects with different levels of difficulty, that is, the recognition and recall knowledge of each vocabulary component.

The findings from this model also bring to mind the potential misleading effect of dividing a component into its recognition and recall mastery. Although they are two types of mastery of one component, is important to remember that the recall masteries

[20] As in any SEM analysis, the best-fitting model is only one of the possible representations of relationships, and thus the revised unidimensional model is probably not the only valid statistical representation of vocabulary knowledge. Nevertheless, it is the conceptualisation that best fits my data with its particular measures and participants.

of different components are more related in difficulty to each other than to their recognition counterparts, based on the implicational scale reported in Chapter 4. That is, the form–meaning recognition and form–meaning recall aspects are just different difficulty levels of the single vocabulary dimension, in which the recognition components seem to share more with each other than with their recall counterparts. For example, the form–meaning recognition knowledge of the word 'close' would be more related to the derivative recognition knowledge of 'closed' as an adjective than to the recall mastery of the form–meaning link of such word. Thus, combining the results from the IS and SEM analyses, it appears that the recognition/receptive vs. recall/productive knowledge is the key distinction in vocabulary knowledge, rather than the distinction between the components. This is in line with some theories describing the distinction between receptive and productive knowledge as a different dimension of vocabulary (Henriksen, 1999). That is, the ability to recognise, for example, collocations is very different (in terms of difficulty) from the ability to produce them. They are both part of the same underlying notion, but the degree to which they are known is different. Therefore, according to this empirically-tested conceptualisation of vocabulary, we cannot speak of collocational knowledge overall, but rather recognition and/or recall knowledge of collocations.

6.6.4. Contribution of components to overall vocabulary knowledge

Research Question 1 concerned the contribution of the different components to vocabulary knowledge. The results show that these individual aspects contribute to explaining vocabulary knowledge almost at the same level. The high factor loadings (.81-.93) of the various word knowledge aspects to vocabulary knowledge are another demonstration of the strong intercorrelations among all the aspects through the vocabulary knowledge construct. All these components combined explain 66% to 87% of the variance in vocabulary, leaving no doubt to their essentiality in representing vocabulary knowledge and their strong contribution in explaining the construct.

Collocation recognition seems to be the aspect that explains the least amount of variance, although with minimal difference. We could consider this small variation as another indicator of the special status of collocational knowledge in vocabulary. Although it is generally conceptualised as depth, it has been suggested that is not

completely clear whether this aspect is closer to representing vocabulary size or depth (Gyllstad, 2013). Collocations are, hence, special types of lexical units, and as such they can also be assessed on their form and meaning (e.g., Peters, 2016; Webb et al., 2013). Thus, some have claimed that they should be included as representations of the vocabulary size of a learner (Cobb, 2013). Due to this singular nature, the strength of contribution of collocations to overall vocabulary knowledge seems slightly poorer than that of aspects such as recall of derivatives (87%).

Derivational knowledge, both recognition and recall, came out as the best descriptors of vocabulary knowledge, explaining 85% and 87% of the variance respectively. This is consistent with multiple studies that include morphological knowledge/awareness as one of the best indicators of vocabulary knowledge (Kieffer & Lesaux, 2012b; McBride-Chang et al., 2008; Zhang & Koda, 2011; Zhang, 2016). Multiple-meanings knowledge follows in contribution of variance with 80% for both recognition and recall. This aspect had also been shown to be the strongest contributor to vocabulary knowledge in previous studies (Tannenbaum et al., 2006). My findings indicate that the intrinsic difficulty of derivatives and multiple-meanings aspects, as shown in Study 1 (Chapter 4), makes these components better indications of the overall vocabulary knowledge of a person. In other words, measuring these more complex aspects can provide a better representation of general vocabulary knowledge than easier aspects.

Finally, the form–meaning link explains 72% of the variance in recall and 75% in recognition. It may have been expected that the form–meaning link would have a stronger relationship to vocabulary knowledge measurement than any other aspect, due to its importance for basic communication (Laufer & Paribakht, 1998). However, this result seems to indicate that form–meaning link, being one of the easiest aspects to achieve for learners, means it is not the best indicator of overall vocabulary knowledge, since its knowledge does not imply knowing as many types of knowledge as derivative or multiple meanings do. This is also clearly connected to my finding that recall aspects tend to contribute to vocabulary knowledge slightly more than their corresponding recognition aspects. This larger contribution of recall aspects and weaker influence of recognition ones was also found by Ma (2015), and are in line with Schmitt's (2014) conclusion that more complex, deeper aspects interrelate with each other more strongly than with easier aspects.

Overall, the findings suggest that all the word knowledge aspects are essential and equally important to achieve comprehensive knowledge of vocabulary. However, this does not indicate that these aspects of vocabulary are equal in difficulty. Relating the findings in this study to those from the implicational scale in Study 1 (Chapter 4), vocabulary knowledge is a unique dimension comprised of various aspects of word knowledge which are very highly interrelated but differ in their difficulty level for L2 learners. Thus, although there are no boundaries between these aspects of knowledge in the mental lexicon, they still pose different burdens to L2 learners, and thus would be expected to be learnt at different stages. As a consequence, dividing the aspects as separable components may still be a desirable and useful solution for the purpose of teaching and learning. This would raise learners' explicit awareness of the different aspects involved in knowing vocabulary, as well as helping instructors to organise the teaching according to their difficulty level.

To conclude, the unidimensional description of vocabulary knowledge as comprised of the combination of the independent recognition and recall knowledge of every aspect seems to be the best statistical representation of vocabulary knowledge. This conceptualisation suggests that for learners to achieve good mastery of vocabulary knowledge, the full range of word knowledge types need to be enhanced. In the same way that words are all linked in multiple ways, so that learning one word has some effect on learning others (Meara & Wolter, 2004), my findings suggest that the word knowledge aspects are not unrelated and independent from each other, but rather knowledge of one aspect influences the acquisition of others.

6.6.5. Consideration of terminology

The lack of empirical support for the multidimensionality of vocabulary knowledge does not indicate that Nation's framework of word knowledge components is erroneous. In fact, it seems logical and practical in terms of pedagogy to make distinctions between the various word knowledge components. However, my results raise a red flag regarding the terminology employed to describe them, and suggest that naming the word knowledge components *dimensions* or *constructs* is not appropriate.

A *construct* is defined as "a meaningful interpretation of observed behaviour" (Chapelle, 1998, p.33). It refers to an abstract concept of some type of feature, attribute,

ability or knowledge that has two characteristics: 1) it can be related to other concepts or constructs through hypotheses, and 2) it can be observed and measured through specific instruments (Abreu, 2012). However, constructs are latent, unobservable properties, and thus they cannot be measured directly, but only through the participants' performance on particular observable measures. Furthermore, *dimensions* are subordinates of a construct, that is, the sub-constructs that comprise a larger construct. As such, both concepts share the same properties.

The terms *dimension* and *construct* have been employed inconsistently in the vocabulary literature. They have been used as both synonyms, or in a hierarchical relationship, so that for some authors *construct* is a high-level concept (e.g., Henriksen, 1999) and for others *dimension* is the superordinate, which comprises multiple constructs within it (e.g., Gyllstad, 2013). Within the components approach, researchers have generally understood vocabulary as the general construct, and called the different components *dimensions* of vocabulary knowledge (Henriksen, 1999; Milton & Fitzpatrick, 2014; Read, 2000). However, in statistics, the use of the terms *dimension* and *construct* have the same implications: if two attributes of knowledge are considered to be different dimensions, then they are essentially different from each other and independent in nature, although they can still be correlated. According to my findings, what most conceptualisations of vocabulary knowledge call *dimensions* of vocabulary knowledge, are not so in real terms, since they cannot be empirically discerned from the other aspects they are related to.

As a result, keeping the term *dimension* to describe the word knowledge components is problematic and can be misleading, since it implies that the components can still be treated theoretically as discernible latent concepts, when research seems to suggest they are not. Hence, I propose the term *dimension* not to be employed to label the various word knowledge components, and to keep it to define vocabulary knowledge overall. In turn, vocabulary knowledge can also be described in general as a *construct*, in the sense that it is essentially distinct from other abstract concepts (such as grammar), but understanding that the construct of vocabulary knowledge as a whole comprises only one dimension. Although this terminology choice may just seem harmless, it is important to bear in mind that the use of terms carries some consequences that can affect both future theory and research. If components keep being addressed as dimensions, future research might consider these components as fundamentally different concepts.

Therefore, when reporting research on vocabulary knowledge, it is essential to be precise and accurate in the use of such terms and what they refer to, since they have both theoretical and practical implications.

6.6.6. Limitations and directions for future research

As in any study, there are some limitations that could be the stepping stone for future research. In the first place, vocabulary knowledge has been conceptualised in many different ways. This study only investigates one of those conceptualisations (i.e., Nation's 2013). The different ways in which vocabulary is understood would influence the way it is measured, and thus it is possible that these other ways of looking at vocabulary would lead to different conclusions (e.g., Kieffer & Lesaux, 2012a). A question for future research is whether other hypotheses and other measures of vocabulary knowledge would confirm the unidimensionality of vocabulary or allow the different components to be separately identified in a multidimensional model.

Moreover, the vocabulary knowledge construct as described by Nation (2013) comprises more components than the ones included in this study. It would be interesting for future research to test other components of word knowledge and compare whether the model of relationships change when these aspects are included. Specially, components such as constraints on use or grammatical functions have been rarely investigated, in particular in multicomponent studies (Schmitt, 2010). Thus it is unclear whether their intercorrelations with other components are as high, and thus, whether the unidimensional model would still hold in this case. Research exploring this issue would provide a more comprehensive picture of the overall nature of vocabulary knowledge.

The cross-sectional nature of this study also limits the implications from the findings. Because the data was not collected in a longitudinal manner, no claims on the developmental changes in vocabulary according to proficiency can be made. For example, two studies claiming multidimensionality of vocabulary (i.e., Kieffer & Lesaux, 2012a; Scott et al., 2017) examined this knowledge on children living in the L2 environment. Thus, it is possible that vocabulary knowledge behaves in a more multidimensional manner for a specific proficiency group than other. A focus of future research should include a longitudinal examination of vocabulary components across

proficiency levels, to see whether the evolution in the different components provide any different results.

Another limitation is the examination of vocabulary knowledge by speakers of only one L1. This leaves us unsure of whether this conceptualisation of vocabulary knowledge as a unidimensional model can be applied to any other EFL learners. It has been widely recommended to conduct more replication research on the stability of results across samples, especially in studies with learners of varying characteristics (In'nami & Koizumi, 2011). Specifically, since Study 2 in Chapter 5 showed that cognateness had an effect on the knowledge of most vocabulary components, it is possible that for learners of English from a non-cognate L1 the relationships between word knowledge components differ. In order to obtain a more complete understanding of the nature of vocabulary in English as an L2, this research should be replicated with a non-cognate language, and preferably a larger sample. Chapter 8 presents this replication.

6.7. Conclusion

The current study examined the nature of vocabulary knowledge following Nation's (2013) conceptualisation of word knowledge as a multidimensional construct. SEM was employed to empirically test the validity of this multidimensionality of vocabulary as well as the relationships between these dimensions. The findings provide empirical evidence that all the components previously described as belonging to word knowledge do indeed form part of this construct. However, the results also seem to indicate that the word knowledge components do not behave as independent concepts in the mental lexicon, and therefore, they cannot be theoretically understood as independent dimensions. Rather, they all are very highly interrelated parts of the same unique vocabulary knowledge construct. This unidimensional description of vocabulary seems to be the best empirically-supported conceptualisation of vocabulary knowledge to date. Thus, this study provides the initial step towards the development of an empirically-based general theory of vocabulary acquisition.

Chapter 7

Order of acquisition of vocabulary knowledge components in L2 English: A replication study

Study 1 discussed in Chapter 4 showed that a reliable order of acquisition of vocabulary knowledge components exists for Spanish EFL learners. Although this order proved to be robust under different accuracy criteria, claims cannot be made about the transferability of such order to other learner populations. Therefore, the question remains whether this order of acquisition is language dependent or holds for EFL learners of different language backgrounds. In the scenario that the order remained constant for different populations, an argument could be made about a generalisable scale of acquisition of word knowledge components. This chapter reports on a replication study of the knowledge and order of difficulty of word knowledge components by Chinese learners of English.

7.1. Background of the study

Chapter 4 discussed the need for the investigation of how vocabulary knowledge is acquired by second or foreign language learners, and presented the methodology required for that investigation. It also reported on a study that establishes a valid and systematic order of acquisition of word knowledge components by Spanish EFL learners. Yet, Spanish is a Romance language which shares a considerable number of cognate words with English. The influence that cognateness status of words has on the knowledge of the different vocabulary components was reported in Study 2 (Chapter 5), which revealed a strong effect of cognateness on most of the word knowledge components. In view of the findings in Studies 1 and 2, a question arose regarding the effect that the cognate nature of the participants' mother tongue had had on the

implicational scale. In order to answer this question, Study 1 in Chapter 4 has to be replicated on an EFL learner population from a non-cognate background.

Replications of studies are essential in scientific research in order to test the robustness, validity and generalisability of previous results. Yet, they are still very uncommon. Recent reviews have shown that replication studies are especially scarce in the field of SLA research, as compared to disciplines such as psychology and education, despite the calls for replication in this field (e.g., Porte, 2012; Porte & Richards, 2012). Marsden, Morgan-Short, Thompson, and Abugaber (2018) report that only one in every 400 journal articles in SLA are replicated. Thus, undertaking replications in language learning is deemed a necessary route in SLA research. For this reason, in order to test the validity and generalisability of the order of acquisition of vocabulary components, I consider that a systematic replication study is essential for a more accurate interpretation of the results. Systematic replications repeat the research design almost exactly, but involve new data from new participants, allowing for the comparability between the original and replication studies and the validity of the phenomenon measured (Porte, 2012). Thus, the purpose of this type of replication is to test for the generalisability of results, making it the most appropriate for the purpose of my study. If the scale order proves to be different across both populations, it could be claimed that the cognate status of Spanish and English affected the order of acquisition of components, and thus, it may be applicable only to Romance-based languages. On the other hand, if the implicational scale proves to remain intact regardless of the participants' L1, it would be an indication that there may exist a universal natural order of vocabulary acquisition. The present study reports on the results of such replication.

To avoid repetition, this study moves straight to the methodology part. Refer to Chapters 4 and 5 for background on implicational scaling analyses and the cognate effect, respectively.

7.2. Purpose and research questions

The current study investigates the knowledge that Chinese learners of L2 English have of the various vocabulary components. The main objective of this study is to try and establish a reliable order of word knowledge acquisition for this population and compare it to the one retrieved for the Spanish sample.

The study aimed to answer the following research questions:

1. What knowledge do Chinese learners have of vocabulary components and how is it different from that of Spanish learners?
2. Can an order of acquisition of word knowledge components also be established for EFL learners of non-Romance languages?
3. If so, how does this scale differ from the one found for Spanish speakers?
4. Is there a universal order of acquisition of word knowledge components?

Given the findings in Study 1 (Chapter 4), I hypothesise that a systematic order of difficulty in the acquisition of word knowledge aspects can also be found for the Chinese population. However, based on the effect of cognates on general word knowledge, it is also expected that this order would vary in EFL learners of Romance languages and non-Romance languages (see Study 2, Chapter 5).

7.3. Methodology

The main aim of the study is to compare the order of acquisition of vocabulary components of two different language populations of EFL learners. Therefore, the implicational scaling study in Chapter 4 was replicated employing exactly the same methodology and test-battery. For a detailed discussion of the development of materials and a description of the scoring protocols, see Chapter 3. For a complete description of the Guttman scaling procedure and the selection of the two cut-off criteria employed to evaluate mastery of word knowledge components in implicational scaling, refer to Chapter 4, Section 4.3.

7.3.1. Participants

The target population in this study was EFL learners of English of a non-Romance language. In order to discover whether the scale found for Spanish learners in Study 1 (Chapter 4) holds in a different learner population, I aimed for speakers of a language as different as possible from that in Study 1 (Spanish) and from the target L2 (English). The logic behind it was that if the order remains with such different EFL

learner populations, it is possible that a universal order of acquisition of vocabulary components might exist. Thus, the participants of the present study were Chinese EFL learners of English.

In cases in which the participants have the same proficiency level, finding an IS is impossible (Hatch & Lazaraton, 1991). Therefore, as in Study 1 in Chapter 4, I made sure I targeted a sample of Chinese learners with different proficiencies in English, from beginners to advanced. 173 participants took part in the study. All of them were undergraduate or postgraduate students of various degrees in different universities in China and the UK. They were recruited as volunteers with the only requirement of being native speakers of Chinese and having some knowledge of English. Data from three participants was excluded from further analysis due to failing to complete the whole test. The final pool of participants included in the analyses consisted of 170 subjects (152 females (89.4%) and 18 males (10.6%)). Age ranged between 18-29 years old (M = 20.03, SD = 1.93).

The minimum amount of time the participants had studied English in the form of formal instruction or self-instruction was one year, and the maximum was 16 years (M = 10.34, SD = 3.20). In fact, most of the participants were studying English formally at the time of the study (87.1%, 148 people) for diverse reasons: 1) to obtain an official certificate to study abroad (e.g., CAE, TOEFL) (27%); 2) because English was the language of instruction during their university degrees (55.9%); 3) as part of a foundation year at university in order to start their degrees (17.1%). Only a fifth of the participants (19.4%) reported having a degree in English (e.g., English teacher education, English Philology, etc.) at the time the study was conducted, and 9.4% were working towards getting such degrees. Overall, this suggests that the vast majority of the participants were in formal contact with English at the time of the study.

Regarding their proficiency, more than half of them (62.4%, 106) reported having an intermediate general proficiency level in English, a third (30.6%, 52 people) rated themselves as beginners in English, and only 7.1% (12) considered themselves advanced users of English. Only 14.1% (24) of the participants had spent a minimum of three months in English speaking countries, up to a maximum of 2 years. Finally, almost half (43.5%) of the participants reported speaking at least another second language apart from English, primarily Spanish, German, Cantonese and Japanese. Moreover, as Mandarin Chinese is not strictly a mother tongue, all the participants

spoke Chinese dialects. This suggests that their general linguistic knowledge and learning skills were substantial.

Table 57 summarises the participants' overall exposure to English from different sources and modes. Watching TV seems the most popular means of English input, followed very closely by reading.

Table 57 Chinese participants' exposure to English (per week).

	0 hours	0-1 hours	1-2 hours	2+ hours
Reading	2.9%	23.5%	20.6%	52.9%
Watching TV	2.4%	20.6%	26.5%	50.6%
Listening to Radio/Music	3.5%	32.4%	20%	44.1%
Using social networks to communicate	19.4%	44.1%	18.2%	18.2%
Playing computer/online games	64.1%	22.9%	5.3%	7.6%

From this information we can conclude that, regardless of the participants' proficiency levels, they generally had a reasonable amount of contact with English.

7.3.2. Chinese translation

The measurement instrument for data collection needed to be translated into Chinese to accommodate the new sample. Two of the tests in the text battery (i.e., form–meaning link and collocation tests) presented the target words in a context written in the participants' mother tongue. Moreover, as in Study 1 (Chapter 4), the participants were given the option of completing the test with all the instructions and background questionnaire in English or in their L1. Finally, the consent form and information sheet for participants were also offered to them in those two languages. All these sections were translated into Chinese by a professional English-Chinese translator and PhD student in English, who was paid for this service. She consulted me throughout the process to make sure the right message was conveyed. The translations were double-checked and compared to their English version by a separate Chinese L1 speaker with a near-native English proficiency.

7.3.3. Piloting

In order to make sure that the test battery functioned well for the Chinese population, the test was piloted with four Chinese speakers living in the UK and two Chinese speakers proficient in English and living in China. Since this study is a replication, the piloting focused on identifying any issues related to four aspects: 1) language/translation, 2) unclear explanations or sections, 3) particularly difficult items, and 4) cross-influence from test to test. No participant, either in the piloting or the main study, reported any misunderstanding due to a problematic translation or language-related issues. No problems were identified regarding the clarity of instructions or how to complete the test. Nevertheless, some participants recommended presenting examples of the recall multiple-meanings test orally when administering the battery. This suggestion was taken into account. As for the item difficulty, participants reported that, as expected, lower-frequency words would be generally unknown. Finally, post-piloting interviews were conducted to see if participants were able to use information from previous tests on later tests. Participants commented that, although they could remember having seen the target words before, they could not remember any information that helped them answer subsequent tests.

Based on the results from the piloting study, the test battery was considered to work well for the Chinese population.

7.3.4. Test-battery administration

The data collection took place in the UK and China for 2 months, and followed the same process as in Study 1 (see Chapter 4, Section 4.3.4 for more details). The administration of the test was also in pen-and-paper format, and the researcher collected each individual test after having been completed by each participant. After introducing the study orally, the participants read the information sheet and signed the informed consent form (see Appendix 10 for the Participant Information sheet and a sample of signed consent forms, which were provided in English or Chinese based on the participants' preferences). The language background questionnaire was filled in after the whole test battery was completed. The whole test took about 2.5-3 hours to complete. All the participants were given gifts for the value of £8 for their participation,

and a detailed report of their results in each independent test was provided to those participants who requested it, in exchange for their help.

Despite having the option to take the test in two sessions (see Chapter 3, Section 3.3.3), all participants completed it in one sitting. They were tested individually or in small groups according to their availability. Participants were given the option to choose from the test with English or Chinese instructions. Following the counter-balanced designed of the test, participants were randomly assigned to one of the two versions of the test (see Chapter 3 for more details). 86 subjects took Version 1 of the test (*N*=71 Chinese Version 1; *N*=15 English Version 1), while 84 completed Version 2 of the test (*N*=72 Chinese Version 2; *N*=12 English Version 2).

7.3.5. Data analysis

To check the hypothesis that a stable pattern of development in the acquisition of vocabulary can be found for the Chinese population, the main method of analysis employed was implicational scaling (see Chapter 4, Section 4.1.1 for a complete description). Implicational scaling has proved to be useful for the establishment of reliable orders of acquisition and developmental sequences in areas such as grammar and phonology, and Study 1 (Chapter 4) demonstrated that it is also a valuable instrument for the exploration of the order of acquisition of depth of vocabulary knowledge. As expressed in Chapter 4, a general implicational scale of vocabulary acquisition would imply that, if a particular word knowledge aspect is present in someone's vocabulary (e.g., recall of form–meaning link), then an easier component would also be (e.g., recognition of form–meaning link). An advantage of implicational scaling, as compared to providing only mean scores, is that it allows us to look at the individual's performance and establish individual as well as general patterns. The data analysis in this study followed the same process explained in Chapter 4, Section 4.3.5 (refer to this section for details).

7.4. Results and interpretation

The results will be presented in two sections: general results and implicational scaling results. The scoring protocols for the different tests in the battery followed the

same procedure as in Study 1 (see Chapter 3, Section 3.2 for details). As in Study 1, the data for the present study was analysed using SPSS version 24 (IBM Corp., 2016) for general statistics and MSPWin5 (Molenaar & Sijtsma, 2000) for Mokken analysis.

7.4.1. General results

7.4.1.1. Reliability of the test battery

The reliability of each individual test in the battery was checked by means of the Cronbach's Alpha reliability value (number of participants $N = 170$; number of items in each test $k = 20$). Table 58 shows that the reliability levels were good for half of the tests (i.e., derivatives and multiple-meanings tests) and slightly under the acceptable threshold of $\alpha = .70$ for the other half (i.e., form–meaning and collocations).

Table 58 Internal consistency reliability for word knowledge aspects' tests (Chinese group).

Test	Cronbach Alpha (α)
Form–Meaning link Recall	α = .60
Form–Meaning link Recognition	α = .50
Derivative Recall	α = .83
Derivative Recognition	α = .81
Multiple-Meanings Recall	α = .81
Multiple-Meanings Recognition	α = .81
Collocation Recall	α = .67
Collocation Recognition	α = .60

Cronbach alpha assesses how well the items in a scale are related to each other, that is, measure the same concept. While it is generally accepted that Cronbach alpha should be over .70, this metric is sensitive to multiple conditions, such as the number of items in the scale and the intercorrelations of items (DeVellis, 2012). For example, it has been suggested that an alpha calculated on Pearson correlations for dichotomous items will be an underestimate of the real internal consistency of a test (Brooks, 2014). This is the case for most vocabulary tests, and for the four tasks with lower alpha values in my study. As a result, it has been suggested that alpha values of around .60 are

probably acceptable for vocabulary studies (Spencer et al., 2015; Zhang & Yang, 2016). Yet, Cronbach alpha is probably not the best reliability metric for vocabulary measures, as each word is its own concept (i.e., knowing a word does not imply knowing another one). Hence, the sophisticated reliability analysis of Average Variance Extracted and Composite Reliability performed in Chapter 8 (see Section 8.4.2) are more appropriate measures for vocabulary tasks, and were the ones employed to decide on the reliability of the tests. Since these measures yielded good reliability values for each task, the tests were considered sound for analysis.

The Multiple-Meaning Recall test showed good internal consistency, but it was given to two raters to score, as it required open-ended answers and is amenable to possible inconsistency in the scoring. The two raters were proficient in both English and Chinese (i.e., Chinese native speakers with a postgraduate degree in English), and were trained and given detail scoring guides to score a random 20% sample (n= 30) of the total number of multiple-meanings recall tests analysed. Each test was assessed by each of the two raters.

Intra-class correlation coefficient (ICC) showed a high degree of reliability between the two raters. The ICC across the two inter-raters was .96, with a 95% confidence interval from .91 to .98 (F(29,29)=44.123, p<.001). Therefore, the inter-rater ICC reliability for this test showed a general agreement of 96% between the two raters. This indicates that the responses to this test were reliably scored.

7.4.1.2. Normality check

Kolmogorov-Smirnov test with Lilliefors significance correction was used to check for normality of the distributions of the different tests' scores. These distributions were also visually inspected by means of histograms, and Skewness and Kurtosis values were checked to further decide on the non-normality of the data. Three tests showed to be normally distributed: the VLT Total (D(170) = .042, p=.200), the Derivatives Recall test (D(170) = .060, p=.200), and the Multiple-Meanings Recognition test (D(170) = .049, p=.200). The rest of the tests' scores deviated from normality (Form–Meaning Recall, D(170) = .103, p=.000; Form–Meaning Recognition, D(170) = .134, p=.000; Derivative Recognition D(170) = .072, p=.031; Multiple-Meanings Recall D(170) = .089, p=.002; Collocation Recall D(170) = .090, p=.002; Collocation Recognition

D(170) = .160, *p*=.000). Therefore, non-parametric tests were employed for overall analysis.

7.4.1.3. Knowledge of the vocabulary components

Descriptive statistics were computed to obtain an estimation of knowledge in the different word knowledge components from the whole sample. Table 59 shows the raw scores for each individual test.

Table 59 Descriptive statistics of the Chinese participants' performance (raw total scores).

Measure	Min.	Max.	Max. Possible	Mean	*SD*
VLT TOTAL	30	116	120	74.01	16.11
VLT 2K	16	30	30	27.73	2.44
VLT 3K	9	30	30	22.19	5.43
VLT 5K	0	30	30	19.17	5.91
VLT 10K	0	26	30	4.92	5.30
Form–meaning Recall	2	17	20	10.63	2.40
Form–meaning Recognition	10	20	20	15.88	1.91
Derivative Recall	23	67	80	41.26	8.21
Derivative Recognition	22	71	80	52.66	8.59
Multiple-Meanings Recall	15	48	60	30.18	5.99
Multiple-Meanings Recognition	17	55	60	38.94	6.97
Collocation Recall	4	20	20	11.44	3.08
Collocation Recognition	8	20	20	15.46	2.50

Note: *N* = 170

Since each word knowledge component had a different scoring system and maximum score, to make the information more interpretable, the descriptive statistics are presented in percentages in Table 60.

Table 60 Descriptive statistics of the Chinese participants' performance (percentage total scores).

Measures	Min. %	Max. %	Mean %	*SD*
VLT TOTAL	25	96.67	61.67	13.43
VLT 2K	53.33	100	92.43	8.12
VLT 3K	30	100	73.96	18.11
VLT 5K	0	100	63.90	19.70
VLT 10K	0	86.67	16.39	17.64
Form–meaning Recall	10	85	53.15	11.99
Form–meaning Recognition	50	100	79.38	9.53
Derivative Recall	28.75	83.75	51.57	10.26
Derivative Recognition	27.50	88.75	65.82	10.74
Multiple-Meanings Recall	25	80	50.29	10.00
Multiple-Meanings Recognition	28.33	91.67	64.90	11.62
Collocation Recall	20	100	57.21	15.38
Collocation Recognition	40	100	77.32	12.50

Note: $N = 170$

Table 61 shows a broad range in the minimum and maximum test scores, indicating that the target items in each test differed in their difficulty and that the Chinese population was diverse regarding vocabulary knowledge, thus meeting the sample requirements for implicational scaling analysis. The Chinese learners knew, on average, at least half of the items in each test. The recognition knowledge of each aspect by the Chinese learners shows a higher mean score than its recall knowledge. Moreover, the form–meaning recognition aspect showed to be the best known aspect on average also for the Chinese group, while overall knowledge of multiple meanings and derivatives proved to be the most complex.

In the case of derivative knowledge, the Chinese learners were also found to confuse the suffixes required to create the different word classes. For example, some instances were found where learners recalled the suffix *-ly* as an adjective instead of adverb (e.g., *hard* as noun, *hard* as verb, and *hardly* as adjective), or *-en* as adjective instead of verb (e.g., *broad* as noun, *broad* as verb, *broaden* as adjective). In other cases

they knew some suffixes to create adjectives and adverbs, but applied them to the word wrongly (e.g., *bright* as both noun and verb, *brightful* as adjective and *brightfully* as adverb; or *distinction* as noun, *distinct* as verb, *distinctful* as adjective and *distinctfully* as adverb).

Moreover, in the recognition of derivatives learners were found to be prone to selecting a word for each word class instead of contemplating the possibility that some forms did not exist. For example, interpreting *freshness* as noun, *fresh* as verb, *fresh* as adjective, *freshly* as adverb. In some cases, the distractors that followed the morphological rules were chosen as the correct option before considering the non-existence of a particular word class (e.g., *bank* as noun, *bankate* as verb, *bankful* as adjective and *bankedly* as adverb, or *indent* as noun, *indentify* as verb, *indented* as adjective and *indentively* as adverb). In other cases, the suffixes chosen were appropriate for the word class but not for the specific word being tested (e.g., *charmness* as noun, *charm* as verb, *charmful* as adjective and *charmfully* as adverb; or *bright* as noun, *brighten* as verb, *brightful* as adjective and *brightly* as adverb).

These mistakes in the derivatives knowledge suggests that the distributional derivational mastery of the Chinese participants, which controls the form-function relationship of morphemes, was also incomplete. That is, the learners may know some suffixes and morphological rules used to create the different forms, but sometimes apply them either to the wrong word class or to the wrong word.

Knowledge of multiple meanings was found to be problematic for the Chinese learners as well. Some target words, such as *grate*, were unknown by most learners even at the form–meaning level, and thus completely unknown in the multiple-meanings aspect. Other words were known in one meaning but not others, such as for the word *charm*, which was rarely known with the meaning of 'spell', or *bank* with the meaning of 'tilting'.

The mean score of the collocation recall test showed to be higher than that of the form–meaning recall test also for the Chinese learners. This provides support for the hypothesis that the similarity of the test formats and the diverse frequencies of the words to be recalled had an influence on the difficulty of both tests. It seems that the fact that in both tests learners had to recall the form of a single word given the first letter, along with the fact that the forms to be recalled in each cased varied largely in their frequency

range (i.e., 1K-9K for the form–meaning and 1K-3K for collocations), made the collocation recall test a generally easier task than the form–meaning recall test.

Regarding the VLT Total scores, we see that the Chinese participants knew on average 62% of the words tested (raw score 74.01/120, *SD* = 16.11), which suggests that the overall vocabulary size of these participants was good. However, when we look at the mean percentage score of the individual VLT sections we see dissimilar behaviour between them. The knowledge of the VLT 10K level by the Chinese learners was very limited, answering correctly only 16.40% of the items on average. This indicates that the sample did not have a good mastery of low-frequency words. Their knowledge of the VLT 5K proved to be much better, with a mean percentage score of 64, suggesting that their knowledge of mid-frequency vocabulary was acceptable. Finally, the Chinese learners proved to have considerable mastery of the VLT 3K and 2K, with a mean score of 74% and 92% respectively. Overall, the Chinese participants in this study have mastered the highest frequency words, while their knowledge of vocabulary seems to decrease smoothly as words become less frequent until the 5K level, with a sharp reduction in knowledge of very low-frequency words.

7.4.1.4. Correlations between word knowledge components and the VLT

In order to better understand the extent to which the word knowledge components are interrelated with each other and with the VLT scores, correlational analyses were calculated. Spearman's rho correlations were computed due to the univariate non-normality of the tests' scores. Table 61 shows these correlations.

Table 61 Correlations between word knowledge components (Chinese group).

	VLT Total	FM Recall	FM Recog	Deriv Recall	Deriv Recog	MM Recall	Collo Recall	MM Recog
VLT Total								
FM Recall	.619**							
FM Recog	.567**	.440**						
Deriv Recall	.775**	.483**	.579**					
Deriv Recog	.594**	.439**	.551**	.711**				
MM Recall	.751**	.611**	.570**	.721**	.656**			
Collo Recall	.568**	.454**	.492**	.521**	.533**	.588**		
MM Recog	.613**	.466**	.620**	.602**	.592**	.722**	.550**	
Collo Recog	.612**	.433**	.485**	.561**	.493**	.603**	.701**	.497**

Spearman: $^{**}p < .01$

VLT = Vocabulary Levels Test; FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.

It can be seen that the correlations between the word knowledge components for the Chinese group are all significant at the p <.01 level and lie between .43 and .78, indicating medium to strong interconnections (Cohen, 1988; Plonsky & Oswald, 2014). This suggests that there is a relatively strong tendency that knowing an aspect is related to better knowledge in the other aspects. Moreover, knowledge of the various vocabulary components correlate positively and very strongly with the participants' vocabulary size, as indicated by the VLT Total score (range r =.57 to .78). The highest correlation with the VLT score is the Derivative Recall test r =.78, followed by Multiple-Meanings Recall r =.75. This suggests that the VLT Total score explains more variance in these tests (60% and 56% respectively) than in the other tests. This higher variance explained by the VLT Total score on derivatives recall and multiple-meanings recall may be due to the fact that these word knowledge aspects are the most difficult to acquire, and thus only the most advanced learners with larger vocabulary sizes would mastered these components. The aspects with the lowest correlations with the VLT Test score were the Form–meaning Recall test r =.567, and the Collocation Recall test r =.568, but still show a large interrelation. Moreover, the correlations between the VLT scores and the recall tests are slightly higher than for the recognition tests, suggesting that, as the vocabulary size of learners increases, the recall mastery of aspects improves

slightly more than their recognition mastery. These results seem to support previous research suggesting the strong relationships between depth and size of vocabulary knowledge, as well as within components of depth of vocabulary knowledge (Qian, 1999; Schmitt, 2014).

For each of the four word knowledge components, the correlations between their recognition and recall knowledge show also high interrelations. The Form–Meaning Recognition and Recall correlated at .44, Derivative Recall and Recognition at .71, Multiple-Meanings Recognition and Recall at .72, and Collocation Recall and Recognition at .70. Finally, the recall aspects correlate between .46 and .72, ($p < .01$) and the recognition aspects between .49 and .62 ($p < .01$), suggesting that, although they are different aspects, they are part of the recall and recognition vocabulary knowledge concepts respectively.

Overall, we observed medium to large correlations between the different word knowledge components, as well as between them and the vocabulary size. This indicates that the different types of word knowledge are strongly interrelated with each other in the mental lexicon.

7.4.1.5. Correlations between the individual words' scores and their raw frequency

In order to explore how the frequency of words related to their knowledge in the different aspects, correlational analyses were computed between the total percentage of correct answers per word in each component and these words´ raw lemma frequency in the COCA. Table 62 shows the correlations between the lemma frequency of each of 20 target words and the average score for each target word in the eight word knowledge aspects.

Table 62 Correlation between words' lemma frequency (COCA) and knowledge of the different aspects (Chinese group).

	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
Lemma Freq	.603**	.415	.839**	.733**	.269	.845**	.671**	.678**

Table 62 shows that the lemma frequency of the 20 words correlated significantly and strongly with the total of correct responses for all word knowledge components except Collocation Recognition or Recall. The reason for this lack of correlation might be that the frequency counts I am using in this analysis refer to only the target word, rather than the whole collocation. This result suggests that the collocate frequency, as well as the frequency of the collocation as a whole, may be needed for the effect of frequency to be attested on collocational knowledge. With this in mind, it seems logical that frequency of a target word does not appear to affect collocational knowledge, while showing some influence on the knowledge of the other aspects. Yet, as explained in Study 2 (Chapter 5), this finding also suggests that there is something in the knowledge of collocations (i.e., formulaic language) that makes it differ from other aspects.

7.4.1.6. Summary of general results

To sum up, the general results show that the best known word knowledge component by Chinese learners is the form–meaning link overall, and that the components for which less knowledge was attested were derivatives and multiple meanings. Recognition knowledge is also known better than recall knowledge. Despite these differences in difficulty, the word knowledge components seem to be very highly interrelated with each other, which suggests that general vocabulary knowledge improves as more is known in one aspect.

7.4.2. Implicational scaling results

The current section will present the implicational scaling results for the Chinese population, investigating the order of acquisition of the different word knowledge components as a whole and per word.

The IS of vocabulary components in this study was created employing the data from 170 participants on their knowledge of the 20 target words in the different aspects. An essential issue when applying implicational scaling is deciding when an aspect is considered acquired or known. As in Study 1 (Chapter 4), I employed two cut-off points for the analysis, a stricter one requiring 80% correct answers per aspect and a more

lenient one accepting 75% correct answers as mastery of an aspect. The word knowledge aspects were then sequenced according to these criteria in order to reveal potential implicational orders.

7.4.2.1. 80% accuracy criterion

To create this implicational matrix, the participants were arranged vertically in order of performance, and the eight word knowledge components horizontally, from better known to worse known. Each subject was coded as having mastered (1) or not (o) each of the eight word knowledge components on the basis that they provided at least 80% accurate answers per aspect.

Following Rickford (2002), I first present the matrix with frequencies of correct answers (i.e., ≥79.5% correct answers), before converting them into 1s and 0s. This frequencies matrix provides an accurate description of the knowledge of the different aspects per person (see Appendix 11). After doing the IS calculation with this percentages matrix, I converted the values into 1s and 0s following the data analysis procedures to produce the final implicational matrix (see Appendix 12).

The percentage matrix yielded the same order of word knowledge components as the binomial matrix, except for one change: knowledge of derivatives recognition appeared before multiple-meanings recognition. However, the scalability coefficient for the percentage table did not reach the minimum level to consider the scale valid (C_{scal} was .53, and thus below the minimum level of .60). Given that implicational scaling is a method to analyse dichotomous data, not percentages, I decided to interpret the results of the 1-0 matrix instead of those of the percentage data, as it is more appropriate for the analysis employed here, and is the one that can be checked by Mokken scaling. The order of acquisition of components suggested by the scale is presented in Figure 21.

Figure 21 Order retrieved in the implicational scale for the Chinese group with the 80% accuracy level (from easier to more difficult).

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

In order to find out whether the scale is valid, and if so, how good it is, I computed the Guttman statistics. The coefficients are presented in Table 63.

Table 63 Results of Guttman Procedure with 80% accuracy criterion (Chinese group).

	C_{rep}	MM_{rep}	% improvement	C_{scal}
Whole matrix	**.95**	.85	.10	**.67**
Recognition part of matrix	**.91**	.73	.18	**.67**
Recall part of matrix	**.98**	.97	.01	**.33**

C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % improvement = Percentage of improvement in reproducibility; C_{scal} = Coefficient of scalability.

By convention, C_{rep} must be over .90 before a scale can be considered valid, and C_{scal} above .60 before scalability can be claimed (see Section 4.3.1 in Chapter 4). Table 63 shows that the implicational scale of all word knowledge components based on the 80% accuracy criterion obtained a good fit and can be considered valid. It can also be seen that the whole matrix and the recognition section produced valid scales. However, the recall section of the matrix alone did not reach the minimum C_{scal}. This result may reflect the fact that the proportion of people who achieved 80% mastery of the components in the recall mastery was relatively low. For instance, only one learner answered correctly 80% of the words in their derivative and multiple-meanings recall mastery. Consequently, the sample in the recall section of the scale does not seem large enough for a scale of recall components alone to be claimed. This finding does nothing but emphasize the great difficulty of recall knowledge for L2 learners, particularly regarding derivatives and multiple meanings.

The coefficient of reproducibility refers to the number of responses in the whole matrix that fit the implicational pattern (Gibson, 1998). The matrix as a whole satisfies the minimum reproducibility value of .90 (i.e., C_{rep}=.95). According to this C_{rep}, 95% of the Chinese participants follow the suggested implicational scale, so that if a participant knows one aspect in recall mastery it would mean that around 95% of the time that person would know all the aspects in recognition. The C_{scal} (.67) indicates that we can be confident that an implicational scale exists, and that the pattern of scalability is robust. Thus, we can conclude that the word knowledge components are scalable for

Chinese learners of L2 English, and the implicational relationship between them is significant and strong.

In order to confirm the results from the IS, I ran a Mokken scaling analysis employing the 1-0 matrix for the 80% accuracy criterion. Table 64 shows the results.

Table 64 Mokken scale for 80% accuracy criterion, reliability and dimensionality assessment (Chinese group).

Order	Aspect	Mean	Item H
1.	Form–meaning Recognition	.58	.56
2.	Collocation Recognition	.55	.52
3.	Multiple-Meanings Recognition	.11	.64
4.	Derivative Recognition	.10	.67
5.	Collocation Recall	.09	.56
6.	Form–meaning Recall	.02	.94
7.	Derivative Recall	.01	1
8.	Multiple-Meanings Recall	.01	1

Scale *H* = .62

Rho = .75

The results from the Mokken analysis show that the order of difficulty of the tests is the same as suggested in the IS illustrated in Figure 21, thus providing support for the previous scale. The homogeneity of the scale is strong (Scale *H* =.62), indicating that the aspects are highly intercorrelated and form part of an underlying construct (i.e., vocabulary knowledge). The reliability of this scale is also high (*rho* =.75). Thus, we can speak of a strong, statistically reliable order of difficulty of the word knowledge components. Besides, none of the individual aspects fall below the recommended value of *H* >.30 to be considered unidimensional. In fact, given their *H* values, they could all be considered strongly unidimensional, suggesting that they all are essential items of the same scale.

Once the scale was demonstrated to be valid by mean of IS and Mokken scaling, I compared the implicational order suggested with that of the Spanish participants. The scale showed exactly the same order for both language groups at the 80% accuracy level. This surprising result may be an indication of the generalisability of an order of acquisition of word knowledge components by L2 learners of English. This order

suggests that all the recognition aspects seem to be easier to acquire than any recall mastery, providing confirmation for the generally agreed notion that recognition knowledge is acquired before recall knowledge, and that moving from recognition to recall mastery in a language is a complex step. The IS also shows that form–meaning link seems to be acquired earlier, and that, when compared to aspects such as multiple meanings and derivatives, knowledge of collocations is not the most difficult aspect for learners.

Overall, applying the 80% accuracy criterion to the results of Chinese learners produces an IS of the various word knowledge components that is statistically valid, and presents an order of difficulty that is consistent across two different EFL learner populations.

7.4.2.2. 75% accuracy criterion

In order to test that the previous results are not just an artefact of the accuracy level selected, I applied a more lenient criterion at the 75% accuracy level. Appendices 13 and 14 show the scalograms for the word knowledge aspects following the 75% cut-off point in percentages and converted into 1-0s, respectively. Both matrices retrieved the same scale. However, the table with percentage information did not yielded the minimum scalability requirements (C_{scal} = .47), and thus, the 1-0 matrix is discussed. Figure 22 shows the order of difficulty of word knowledge components employing this criterion.

Figure 22 Order retrieved in the implicational scale for the Chinese group with the 75% accuracy level (from easier to more difficult).

Form–Meaning Recognition > Collocation Recognition > Derivative Recognition > Multiple-Meanings Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

It can be seen that the pattern of difficulty following the 75% correct answers cut-off point shows practically the same scale as with the 80% accuracy criterion. However, in the recognition section the order of Derivative Recognition and Multiple-Meanings Recognition is swapped with respect to the 80% scale (see Figure 21). This

change indicates that these two components seem to be similar in difficulty for learners, and thus their order may vary slightly depending on the accuracy criterion employed. Inspection of the scalogram in Appendix 14 shows that, indeed, the Derivative Recognition and Multiple-Meanings Recognition tests account for the same number of correct answers (i.e., 35), although the latter test accounted for two more errors than the former (i.e., 18 vs. 16), leading to the order specified in Figure 22. Therefore, this finding suggests that the difference in difficulty between these two aspects is very similar and can be exchanged for this group.

The rest of the scale follows exactly the same pattern as the 80% accuracy criterion, thus supporting the results regarding the order of acquisition of components. Guttman statistics were performed in order to find out whether the scale continues to be valid and reliable with the 75% cut-off point. The coefficients are presented in Table 65.

Table 65 Results of Guttman Procedure with 75% accuracy criterion (Chinese group).

	C_{rep}	MM_{rep}	% improvement	C_{scal}
Whole matrix	**.94**	.85	.09	**.60**
Recognition part of matrix	**.92**	.76	.16	**.67**
Recall part of matrix	**.96**	.93	.03	**.43**

C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % improvement = Percentage of improvement in reproducibility; C_{scal} = Coefficient of scalability.

Again, the overall scale satisfies the minimum C_{rep} and C_{scal} values to be considered valid and robust. As it happened in the 80% accuracy level, both the whole matrix and the recognition section of the matrix met the goodness-of-fit requirements, while the recall section alone did not meet the minimum C_{scal} of .60. This suggests that achieving 75% correct answers in the recall components was difficult for the Chinese participants, and thus the consistency of this section alone is less clear. Table 65 shows that the scalability and reproducibility values of the whole IS are less strong than for the 80% accuracy scale, even though the 75% accuracy level is more lenient. This suggests that, although both scales are valid, the order suggested in the 80% accuracy scale is slightly more robust, and thus, it is to be considered the most consistent. For this reason, and since 80% accuracy is a more appropriate criterion to consider mastery

of a vocabulary component, I will interpret the 80% accuracy scale for the Chinese group as the best representation of an order of difficulty of vocabulary knowledge aspects.

The Mokken analysis for the data with the 75% accuracy criterion yielded the same order as the IS in Figure 22, and produced also good results in terms of the homogeneity and reliability of the scale (see Table 66).

Table 66 Mokken scale for 75% accuracy criterion, reliability and dimensionality assessment (Chinese group).

Order	Aspect	Mean	Item H
1.	Form–Meaning Recognition	.78	.73
2.	Collocation Recognition	.69	.71
3.	Derivative Recognition	.21	.57
4.	Multiple-Meanings Recognition	.21	.62
5.	Collocation Recall	.18	.55
6.	Form–meaning Recall	.06	.75
7.	Derivative Recall	.02	.73
8.	Multiple-Meanings Recall	.02	.83

Scale *H* = .65

Rho = .79

Finally, since it is likely that the difference in difficulty between contiguous aspects varies, I investigated the total number of participants that mastered each aspect across both accuracy levels (see Table 67).

Table 67 Number of participants mastering each aspect (Chinese group).

	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
N of participants mastering 80%	99	94	19	17	15	4	1	1
N of participants mastering 75%	132	118	35	35	31	10	3	3

Table 67 shows that a similar number of learners knew the Form–Meaning Recognition aspect and the Collocation Recognition aspect. Understanding that the degree of difficulty between two contiguous aspects is inversely proportional to the number of participants that mastered each aspect (Abdi, 2010), the results show that the Form–Meaning and Collocation Recognition tests do not differ much in their difficulty, particularly at the 80% accuracy level. Conversely, the jump between mastering the Collocation Recognition aspect and the Multiple-Meanings Recognition one is bigger, suggesting that the latter is clearly more difficult. As it was suggested in the IS for the 75% accuracy criterion, Multiple-Meanings Recognition and Derivative Recognition do not seem to differ much in their level of difficulty. Surprisingly, Form–Meaning Recall seems noticeably more difficult than Collocation recall, particularly at the 75% accuracy level. Derivative Recall and Form–meaning Recall are less similar in difficulty at the 75% accuracy level than at the 80% accuracy level. Finally, recall knowledge of Derivatives and Multiple-Meanings appears to be equally difficult for Chinese learners, with only 1 and 3 people mastering their knowledge.

Overall, it can be concluded that the acquisition of the different word knowledge components along the IS goes through different degrees of difficulty. That is, the order of acquisition seems robust, although the difference of difficulty in moving from one to the next one varies.

7.4.2.3. Implicational scaling per word

In order to explore whether an IS could also be found for each individual word, I applied implicational scaling to each target item. As in Study 1 (Chapter 4), the target words are expected to follow different orders of difficulty of the aspects, as they were selected so that they varied in difficulty within each aspect.

Scalograms for each of the 20 target words were computed, arranging the order of aspects horizontally from easier to more difficult and the participants vertically from better to worse performance. Table 68 presents the order of difficulty of the eight word knowledge components for each target word, and the Guttman scaling coefficients.

The results show that the target words do not seem to follow a single order of difficulty of word knowledge components. Each word presents a different pattern, and these patterns do not follow a logical order of difficulty. However, the Guttman statistics

demonstrate that none of the words' scales achieved the minimum values to be considered acceptable. Although four words reached the minimum coefficient of reproducibility (C_{rep} >.90) (i.e., *close*, *challenge*, *hard* and *indent*), none of the target words reached the minimum scalability threshold (C_{scal} >.60). This indicates that the orders suggested for the individual words cannot be considered valid, and therefore, no IS can be claimed for individual words.

This lack of systematicity in the acquisition of discrete words implies that there are too many individual differences within the words themselves and the participants in order to establish a consistent order of acquisition of vocabulary components for individual words. That is, each word possesses a different difficulty burden of its various word knowledge components, with some aspects being easier in certain words than in others. Moreover, learners' acquire vocabulary in various ways and remember different types of knowledge about a word depending on their individual experiences. The impossibility of finding an IS for the individual words is a reflection of these differences.

Overall, the idiosyncratic nature of words and people means that no IS can be found for particular words, but it can be found for vocabulary knowledge as a whole.

Table 68 Results of Guttman Procedure for individual words (80% accuracy) (Chinese group).

Target word	Order of Aspects	C_{rep}	MM_{rep}	% $_{improve}$	C_{scal}
Close	MMRg>FMRg>CRg>FMRc>CRc>MMRc>DRg>DRc	**.91**	.87	.04	**.31**
Season	FMRg>FMRc>CRg>DRg>MMRg>DRc>MMRc>CRc	**.88**	.80	.08	**.40**
Bank	CRg>FMRg>FMRc>MMRg>CRc>MMRc>DRg>DRc	**.87**	.79	.08	**.38**
Terminal	FMRg>MMRg>CRg>DRg>MMRc>CPc>DPc>FMPc	**.82**	.69	.13	**.42**
Bright	FMRg>CRc>MMRc>MMRg>DRg>FMRc>CRg>DRc	**.88**	.84	.04	**.25**
Distinction	FMRg>CRg>MMRg>DRg>DRc>CRc>MMRc>FMRc	**.81**	.69	.12	**.39**
Draught	CRg>MMRg>FMRg>CRc>DRg>DRc>FMRc>MMRc	**.87**	.76	.11	**.46**
Challenge	FMRg>FMRc>CRg>MMRg>DRg>CRc>MMRc>DRc	**.91**	.87	.04	**.31**
Broad	FMRg>CRg>DRg>FMRc>MMRc>MMRg>CRc>DRg	**.83**	.72	.11	**.39**
Grate	MMRg>CRg>CRc>FMRg>DRg>DRc>MMRc>FMRc	**.87**	.73	.14	**.52**
Mean	DRg>MMRg>FMRg>FMRc>DRc>MMRc>CRg>CRc	**.86**	.78	.08	**.36**
Hard	MMRg>MMRc>FMRg>CRg>CRc>FMRc>DRg>DRc	**.92**	.86	.06	**.43**
Employ	FMRg>CRg>DRg>MMRg>FMRc>CRc>DRc>MMRc	**.85**	.75	.10	**.40**
Fresh	CRg>MMRc>MMRg>CRc>FMRg>FMRc>DRg>DRc	**.87**	.79	.08	**.38**
Charm	FMRg>DRg>FMRc>DRc>MMRg>CRg>CRc>MMRc	**.83**	.66	.17	**.50**
Indent	CRc>CRg>MMRg>FMRg>DRg>DRc>FMRc>MMRc	**.91**	.85	.06	**.40**
Fulfill	FMRg>CRg>CRc>MMRg>DRg>MMRc>FMRc>DRc	**.85**	.76	.09	**.38**
Character	MMRg>CRg>MMRc>FMRg>CRc>FMRc>DRg>DRc	**.89**	.83	.06	**.35**
Redeem	MMRg>CRg>DRg>FMRg>MMRc>CRc>DRc>FMRc	**.86**	.78	.08	**.36**
Development	FMRg>MMRg>CRg>DRg>DRc>FMRc>MMRc>CRc	**.83**	.70	.13	**.43**

For each aspect: *Rg = Recognition; Rc = Recall; C_{rep} = Coefficient of reproducibility; MM_{rep} = Maximum Marginals reproducibility; % $_{improve}$ = Percentage of improvement in reproducibility ; C_{scal} = Coefficient of scalability*
Bold = C_{rep} and C_{scal}; Blue = approaching minimum value; Red = not reaching minimum value.

7.5. Comparison of results between Spanish and Chinese learners

In this section, I explore the similarities and differences in the behaviour of Chinese and Spanish learners regarding vocabulary knowledge. The Spanish learners were on average older than the Chinese learners ($t(312) = 8.19$, $p < .001$), and more female participants were present in the Chinese group ($t(312) = -4.28$, $p < .001$). The Spanish learners had been studying English for more years ($t(312) = 7.91$, $p < .01$), but more people in the Chinese group were still studying at the time the test took place ($t(312) = -4.86$, $p < .001$).

7.5.1. Reliability of tests

A comparison of the tests' reliability scores reveals that Cronbach alpha was considerably higher in all tests for the Spanish group (ranging between .74 and .95) than for the Chinese group (from .50 to .83). These findings confirm Byrne's (2016) claim that measuring instruments are often group-specific in how they function. In particular, this difference in reliability indicates that there was more variation in the scores of the Chinese group than in the Spanish group. In other words, the Spanish participants as a group performed more similarly to each other across the different vocabulary components, specifically in the form–meaning link and collocations tests. These findings seem to reflect variations between the two linguistic communities which led to different performance in the responses to the individual items in each test. This suggests that when employing the same tests for two different L2 learner populations, it needs to be taken into account that the specific characteristics of each group may lead to slightly variable test performance and reliability.

Despite this, the behaviour in the individual tests relative to each other seems to be similar across both populations, in the sense that both groups present the lowest reliability scores in the form–meaning and collocation tests, and the highest in multiple meanings and derivatives. That is, the derivative and multiple-meanings tasks proved to be the most consistent tests across both language populations, and the form–meaning recognition test the least consistent for both groups, although the values were always higher for the Spanish participants.

7.5.2. Correlations among word knowledge components and vocabulary size

The correlations between the word knowledge components for the Chinese group lie between .43 and .78. When these correlations are compared to those for the Spanish participants (see Table 69), we see that they are all smaller: the vocabulary components in the Spanish group correlate between .68 and .95, with most of them being in the .80 value. This finding indicates that the tendency of knowing an aspect better when another is mastered remains across both groups, even though the strength of these relationships seems weaker in the Chinese group.

Table 69 Summary of correlation coefficients between measures, means, and standard deviations as a function of language group.

	VLT Total	FM Recall	FM Recog	Deriv Recall	DerivRecog	MM Recall	MM Recog	Collo Recall	Collo Recog	**Chinese**	*M*	SD
VLT Total	-	.619**	.567**	.775**	.594**	.751**	.613**	.568**	.612**		61.67	13.43
FM Recall	.786**	-	.440**	.483**	.439**	.611**	.466**	.454**	.433**		53.15	11.99
FM Recog	.760**	.773**	-	.579**	.551**	.570**	.620**	.492**	.485**		79.38	9.53
Deriv Recall	.895**	.811**	.803**	-	.711**	.721**	.602**	.521**	.561**		51.57	10.26
Deriv Recog	.865**	.786**	.828**	.945**	-	.656**	.592**	.533**	.493**		65.82	10.74
MM Recall	.872**	.747**	.774**	.845**	.818**	-	.722**	.588**	.603**		50.29	10.00
MM Recog	.826**	.720**	.766**	.835**	.831**	.848**	-	.550**	.497**		64.90	11.62
Collo Recall	.839**	.772**	.749**	.827**	.812**	.817**	.771**	-	.701		57.21	15.38
Collo Recog	.779**	.678**	.700**	.763**	.768**	.752**	.737**	.806**	-		77.32	12.50
Spanish												
M	67.81	53.61	82.81	51.64	61.08	49.69	70.76	60.66	79.69			
SD	16.79	18.91	13.79	20.86	19.90	16.12	15.18	19.05	17.39			

Spearman: **$p < .01$, $N = 144$

Note: intercorrelations for the Spanish group are presented below the diagonal in light-blue shading ($n = 144$) and those for the Chinese group above the diagonal in light-grey shading ($n = 170$).
FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.

Similarly, for both groups, knowledge of the various word knowledge components correlated positively and very strongly with the participants' vocabulary size, as indicated by the VLT Total score. Yet, these correlations were again smaller for the Chinese group than the Spanish group. Despite the difference in the strength of relationships across the two groups, they both show the same pattern: the highest correlation with the VLT score was the Derivative Recall test (Spanish r =.90, Chinese r =.78, p<.01), followed by Multiple-Meanings Recall (Spanish r =.87, Chinese r =.75, p <.01). The aspect with the lowest correlation with the VLT Test score was the Form–meaning Recognition test (Spanish r =.76, Chinese r =.57, p <.01).

Overall, we can conclude that in both language groups the word knowledge components are highly intercorrelated with each other, although the strength of relationships is smaller for the Chinese participants than for the Spanish participants across all aspects. This may indicate that the individual components behave in a slightly more independent manner in the mental lexicon of Chinese learners, although they are still very interrelated.

7.5.3. Correlations between the individual words' scores and their raw frequency

For both the Chinese and Spanish groups, the lemma frequency of the 20 target words correlated significantly and strongly with the total of correct responses for six of the eight word knowledge components (i.e., all except Collocation Recognition and Recall). However, this analysis provides an opposite trend as compared to the previous correlational analyses, in that higher and stronger correlations between the lemma frequency of words and their scores were found for the Chinese group than for the Spanish group in each of the six significant vocabulary components: Form–Meaning Recognition test (Spanish r =.51, Chinese r =.60), Multiple-Meanings Recognition (Spanish r =.66, Chinese r =.84), Derivative Recognition (Spanish r =.48, Chinese r =.73), Form–Meaning Recall (Spanish r =.85, Chinese r =.85), Derivative Recall (Spanish r =.49, Chinese r =.67), and Multiple-Meanings Recall (Spanish r =.52, Chinese r =.68). This finding indicates that the words' lemma frequency influenced Chinese learners' knowledge of each aspect more strongly than Spanish learners' knowledge. That is, the Chinese population seems more affected by the frequency of

words in their vocabulary acquisition process than the Spanish population, who may be more influenced by other aspects, such as cognateness status (see Study 2, Chapter 5).

7.5.4. Knowledge of vocabulary components across groups

The results will be presented in two sections: vocabulary size and depth of vocabulary knowledge.

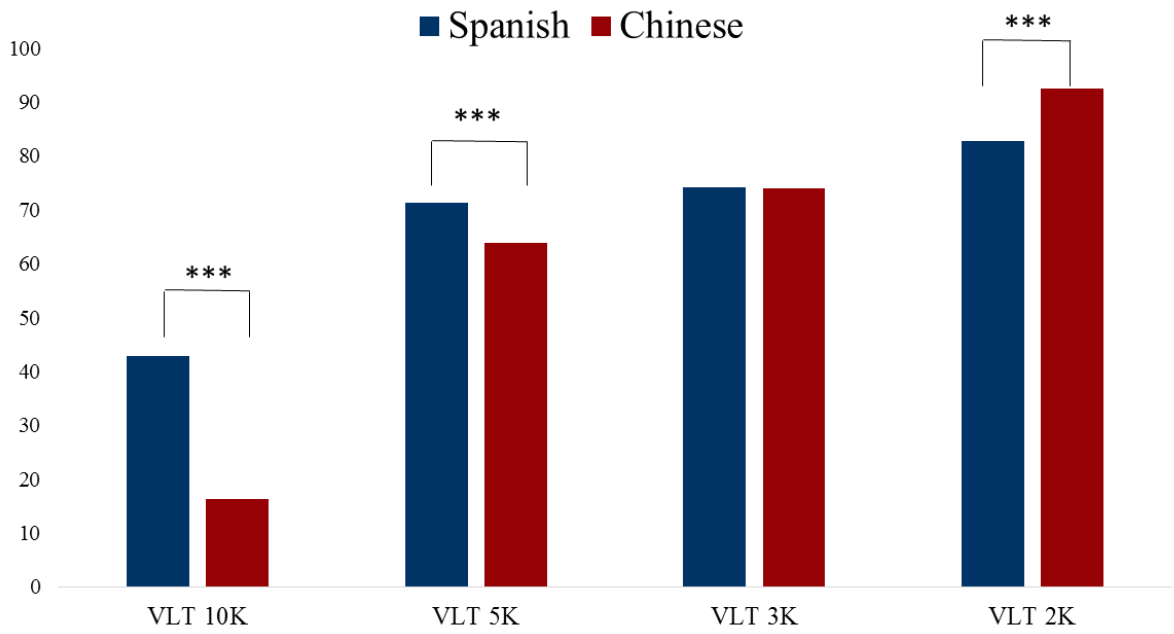
7.5.4.1. Vocabulary size of Chinese and Spanish learners

The VLT Total mean scores of both language groups suggest that the overall vocabulary mastery of the two samples was fairly good overall. The Chinese participants knew on average 62% of the VLT words tested (raw score 74.01/120, *SD*= 16.11), as compared to the 68% of the Spanish group (raw score 81.37, *SD*= 20.15). Although the average total score was similar, differences can be seen between the two groups when looking at the mean percentage score of the individual VLT sections. The VLT 10K mean percentage score for Chinese speakers is 16.40 (*SD*=17.64), whereas the same section for the Spanish participants was 42.84 (*SD*= 18.21). This result seems to support the cognate effect found in Study 2 (Chapter 5), which provide Spanish learners with an advantage on cognate low-frequency words. The scores of the two populations in the VLT 3K is almost the same (Chinese *M* = 73.96, *SD*= 18.11; Spanish *M* = 74.26, *SD*= 20.26), suggesting that both groups perform equally well in the 3K vocabulary level in English. However, the two groups differ in their VLT 5K level, with the Chinese learners performing somewhat worse than the Spanish participants (Chinese *M* = 63.9, *SD*= 19.71; Spanish *M* = 71.25, *SD*= 17.93). On the contrary, the Chinese group performed 10% better in the VLT 2K (Chinese *M* = 92.43, *SD*= 8.12) than the Spanish group (Spanish *M* = 82.87, *SD*= 18.28). This shows that: 1) Chinese learners' knowledge of the 2,000 most frequent words was better than in the Spanish group, and 2) as opposed to the Spanish group, the Chinese group on the whole had mastered the VLT 2K level (28/30 right answers). To sum up, the mean scores in the different VLT sections were lower in the Chinese group, with the exception of the VLT 2K.

In order to statistically compare how the Chinese and Spanish groups differed in their performance in each of the VLT sections, I ran MANOVA analysis (multivariate analysis of variance). MANOVA can analyse several dependent variables simultaneously, taking into account the relationships between these variables, and thus reducing the chances of making Type 1 error. I report the Pillai's trace test statistics for two main reasons: 1) because, in absence of a non-parametric MANOVA test, this test statistics has been found to be relatively robust to violations of multivariate normality, and 2) the two language groups differ in more than one variable, and this statistics has been found to be the most robust in this condition (Field, 2013).

MANOVA test showed that there was a significant difference in the vocabulary size of the two groups, as represented by the VLT tests. $V = 0.65$, $F(12,301) = 46.73$, $p < .001$. Separate ANOVAs exposed significant differences in three of the four VLT levels tested: VLT 10K, $F(1,312) = 170.22$, $p < .001$, VLT 5K, $F(1,312) = 11.77$, $p < .001$, and VLT 2K, $F(1,312) = 37.74$, $p < .001$. However, no significant difference was found between the groups on the knowledge of the VLT 3K level, $F(1,312) = .019$, $p = .89$. Figure 23 illustrates these results.

Figure 23 Comparison of VLT sections' scores by Chinese and Spanish learners.



As it was anticipated by the mean scores, Figure 23 shows that Chinese and Spanish learners differed significantly in their individual VLT scores, suggesting their

vocabulary size was different. The Chinese learners proved to possess a significantly better knowledge of very high-frequency words than the Spanish learners. However, the Spanish learners' mastery of lower-frequency words was substantially better than that of the Chinese learners (probably due to the cognate effect, as discussed in Study 2, Chapter 5).

Although this was not an original research question in this study, these findings are interesting in that the implicational scale shows to remain constant across both L1 populations regardless of their different vocabulary size. This suggests that the scale works not only across different learner populations but also across different proficiencies.

7.5.4.2. Vocabulary depth of Chinese and Spanish learners

Descriptive statistics show that each language group knew, on average, a minimum of half the items in each of the eight word knowledge component. The range of average correct answers per vocabulary component for the Spanish learners was between 49.69% and 82.81%, and for the Chinese group between 50.29% and 79.38%. In both groups the aspect that showed to be the best known on average was the form–meaning recognition, while the worst known aspect was recall knowledge of multiple meanings. Similarly, the recognition knowledge of each aspect showed a higher mean score than its recall counterpart across both language groups.

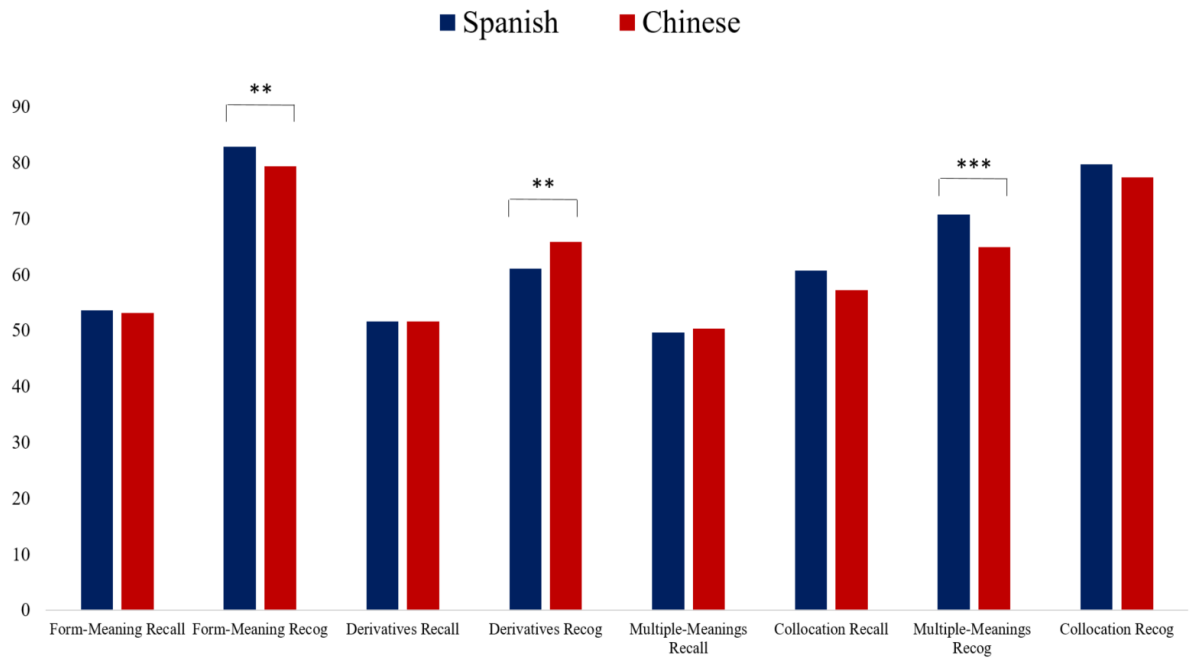
The mean score for each group in each word knowledge aspect shows the similarities and differences in their performance on the specific components. In general, the average results in the recall aspects seem similar in the two groups (Form–Meaning Recall: Chinese *M*=53.15, *SD*=11.99, Spanish *M*=53.54, *SD*=18.82; Derivative Recall: Chinese *M*=51.57, *SD*=10.26, Spanish *M*=51.64, *SD*=20.86; Multiple-Meanings Recall: Chinese *M*=50.29, *SD*=9.99, Spanish *M*=49.67, *SD*=16.12; Collocation Recall: Chinese *M*=57.21, *SD*=15.38, Spanish *M*=60.66, *SD*=19.10). Conversely, the Chinese and Spanish participants tend to differ more in their recognition knowledge, with the Chinese learners performing worse except for derivatives (Form–Meaning Recognition: Chinese *M*=79.38, *SD*=9.53, Spanish *M*=82.81, *SD*=13.79; Derivative Recognition: Chinese *M*=65.82, *SD*=10.74, Spanish *M*=61.08, *SD*=19.90; Multiple-Meanings

Recognition: Chinese *M*=64.90, *SD*=11.62, Spanish *M*=70.76, *SD*=15.18; Collocation Recognition: Chinese *M*=77.32, *SD*=12.50, Spanish *M*=79.69, *SD*=17.40).

These results suggest that the mean score for each individual test differed slightly from group to group, with Chinese learners producing somewhat lower scores except for Derivative Recognition and Multiple-Meanings Recall. These differences in the results were statistically compared by means of MANOVA analyses to test their significance.

Using Pillai's trace, MANOVA showed that there was a significant effect of language group on the performance in the different word knowledge components, $V = 0.65$, $F(12,301) = 46.73$, $p < .001$. Separate univariate ANOVAs on the outcome variables revealed significant differences between the two language groups on their knowledge of three aspects: Form–Meaning Recognition, $F(1,312) = 6.73$, $p < .01$, Derivative Recognition, $F(1,312) = 7.20$, $p < .01$, and Multiple-Meanings Recognition, $F(1,312) = 14.98$, $p < .001$. However, non-significant language group differences were found on the remaining components: Form–Meaning Recall, $F(1,312) = .069$, $p = .79$, on Derivative Recall, $F(1,312) = .001$, $p = .97$, Multiple-Meanings Recall, $F(1,312) = .166$, $p = .68$, Collocation Recall, $F(1,312) = 3.16$, $p = .08$, and Collocation Recognition, $F(1,312) = 1.95$, $p = .16$. Collocation recall was approaching significance. Figure 24 shows the differences across groups in knowledge of the various aspects.

Figure 24 Comparison of scores for each aspect of Chinese and Spanish learners.



MANOVA shows that the two language groups do not seem to differ greatly in their overall test performance. Only three of the eight tests showed a statistically significant difference between knowledge in the two language groups, although the absolute differences for these three tests are not great in real terms. The two language groups show similar results in their recall knowledge, probably due to the large amount of exposure to the English language that both samples reported.

To sum up, the general results suggest a similar pattern in the knowledge of vocabulary components for the Spanish and Chinese groups. The best known word knowledge component by both learner populations is the form–meaning link overall, and the components for which less knowledge was attested were derivatives and multiple meanings. Recognition knowledge is also known better than recall knowledge. Thus, we can conclude that the two groups behave very similarly on the test battery.

7.5.5. Implicational scaling results

This section compares how the two language groups differ regarding the implicational scaling results. For both the Spanish and the Chinese groups, the implicational scale was tested using the raw data as percentages and the data converted into 1s and 0s. For the Spanish participants, the implicational matrix with the percentage scores retrieved the same order of acquisition as the matrix with binomial data (1-0) across both the 80% and 75% accuracy level. In the Chinese group, the percentage scores' matrix and the binomial scores' matrix showed the same scale at the 75% accuracy level, but a slightly different order for the 80% accuracy.

Since implicational scaling was originally designed for the analysis of categorical data, not continuous, the scales derived from the binomial-data matrix yielded better scale fit coefficients across both groups. Thus, these were the scales considered the most appropriate and reliable for interpretation.

7.5.5.1. 80% accuracy criterion

The exact same order of acquisition of word knowledge components was found for the Chinese and Spanish participants under the 80% accuracy criterion (Figure 25). The scalability and reliability of this scale, although slightly higher for the Spanish

group, proved very good for both language populations, suggesting that the scale is to be considered consistent and valid for L2 English learners. This finding provides initial indication of the generalisability of this order.

Figure 25 General order of difficulty of vocabulary components by Chinese and Spanish learners of English (80% accuracy).

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

Research suggests that as the number of linguistic features being studied increases, the difficulty of finding an implicational order decreases, but, if found, the predictive power of the scale becomes more striking (Rickford, 2002). The previous vocabulary studies that applied implicational scaling tested four (Laufer & Goldstein, 2004) and five (Mizrahi, 2016) word knowledge aspects. Thus, the fact that an implicational scale has been found for eight word knowledge components and two different L2 learner populations provides additional power to this order.

7.5.5.2. 75% accuracy criterion

While the results for the 80% accuracy concurred across both language groups, the picture was a bit different for the 75% cut-off level. The Spanish group retained exactly the same scale across both accuracy levels, whereas the Chinese group showed a variation in the 75% level, in which knowledge of Derivative Recognition appeared earlier than knowledge of Multiple-Meanings Recognition. This change suggests that the difference in difficulty between these two components is not very large for the Chinese participants, and therefore, their order can vary under different criteria. However, this 75% accuracy scale had a worse scalability than the 80% accuracy scale for the Chinese group (C_{scal} =.60 vs. .67). What is more, although the scale was statistically valid and reliable, it had the lowest coefficient of scalability across both language groups and scoring criteria, suggesting it is the least reliable scale across both groups. For this reason, I consider the order of acquisition of word knowledge components suggested for the 80% accuracy criterion the most reliable IS. This order

is not only a better estimation of mastery of word knowledge components, it has also been shown to remain intact for three of the four conditions (i.e., across both languages and across both accuracy criteria for one language), making it the most generalisable scale, and thus the best representation of the order of acquisition of word knowledge components in L2's.

7.5.5.3. Extent of difficulty of each component

Table 70 shows some differences between the two groups regarding the extent of difficulty of each aspect (see Table 70). More learners in the Chinese group mastered knowledge of the Form–Meaning Recognition and Collocation Recognition aspects slightly better than in the Spanish group at both accuracy levels. However, for the remaining aspects, Spanish learners noticeably outnumbered the Chinese students in both the 75% and 80% accuracy levels, suggesting that the Spanish group overall had a better mastery of more difficult word knowledge components. This indicates that while the level of difficulty of form–meaning and collocation recognition was more analogous for both groups, the rest of the aspects were considerably more difficult for the Chinese learners than the Spanish learners. Although this difference is evident even from the Multiple-Meanings Recognition aspect onwards, it is the most obvious in the last two components (Derivative and Multiple-Meanings Recall), where only one Chinese learner attained 80% accuracy and three achieved 75% correctness. This indicates that although the two populations of L2 learners follow a regular route on their acquisition process of the different word knowledge components, they develop along this scale at a different pace.

Table 70 Number of participants mastering each aspect (whole sample).

	Accuracy	**FM Recog**	**Collo Recog**	**MM Recog**	**Deriv Recog**	**Collo Recall**	**FM Recall**	**Deriv Recall**	**MM Recall**
					Spanish Group				
N learners mastering	80%	97	91	50	34	32	16	15	7
	75%	107	100	63	50	38	27	25	11

	Accuracy	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
				Chinese Group					
N learners mastering	80%	99	94	19	17	15	4	1	1
	75%	132	118	35	35	31	10	3	3

7.5.5.4. Implicational scale for individual words

The results for the implicational scaling analysis of the individual target words showed the same patterns across both language groups. No single order was found for the various words, and none of the patterns suggested for each of them was found to be statistically valid or reliable. That is, ISs of the acquisition of vocabulary components could not be established for individual words, due to the varied intrinsic characteristics of each item and the participants' individual differences. The conclusion from the implicational scaling results is, therefore, that the order of difficulty of word knowledge components presented in Figure 25 is to be expected when we are measuring general knowledge of an aspect across different L2 learners, but this order does not apply to each word individually.

7.6. Discussion

The present study investigated the knowledge and order of acquisition of vocabulary components of Chinese EFL learners as a replication of Study 1 in Chapter 4. It corroborated certain results from the original study, while at the same time displaying some variances between the two language populations from Romance and non-Romance backgrounds. The two studies in conjunction provide the clearest picture to date of how English vocabulary is acquired in second languages. The discussion will focus on interpreting the implications from the main similarities and differences in the results of both groups.

7.6.1. Knowledge of word knowledge components

Previous multicomponent studies of vocabulary learning have shown that some aspects appear to be better known than others for a particular L2 population. For example, spelling has been found to be one of the better-known aspects, while knowledge of meanings and syntax usually comes later (Chen & Truscott, 2010; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Webb, 2007a,b, 2009). My findings based on two different language learner populations provide support for some of these results, and thus further evidence for the expectation of some vocabulary components being known better than others. In particular, it confirms that the form–meaning link is generally known better, and that L2 English learners struggle with knowledge of derivatives and multiple meanings almost to the same extent.

The best-known aspect in both language groups was recognition of the form–meaning link, which suggests that it is one of the first types of word knowledge that learners master. This finding refutes previous studies showing recognition knowledge of the form–meaning link to be worse known than other components (Chen & Truscott, 2010; Webb, 2007a), while at the same time reinforces the line of research that considers this aspect to be the best-known vocabulary component (Agustín Llach & Moreno Espinosa, 2014; Laufer & Goldstein, 2004; Pellicer-Sánchez & Schmitt, 2010). This seems logical, since this aspect is likely to be the one learners are exposed to the most when learning vocabulary in the L2, as well as the type of knowledge that would enable them to start making use of a language (i.e., recognising the meaning of a word when encountered in reading). Recognition of collocations was confirmed to be easier than recognition of derivatives and multiple meanings, particularly when knowledge of collocates alone was tested. These last two aspects showed to be very problematic for learners when achieving not only recognition vocabulary mastery, but also recall knowledge. Finally, both language groups demonstrated better recall knowledge of collocations than recall of the form–meaning link, confirming the hypothesis that recall of frequent collocates of a target word is easier for L2 learners than form recall of low-frequency target words. This finding provides conditions upon the plethora of research suggesting that L2 learners of English have great difficulty both recognising, but mainly recalling, collocations (Henriksen, 2013; Pellicer-Sánchez, 2017; Wolter & Gyllstad, 2013). In particular, it suggests that learners have a relatively good knowledge of very frequent collocates of words when they are presented in context. My study also

emphasises how the way an aspect is understood and tested affects the estimation of knowledge.

The recognition and recall distinction proves to be fundamental in the learners' vocabulary knowledge. Overall, both groups of participants demonstrated knowing between 60% and 80% of the items in the recognition components, and between 50% and 60% of the recall components. As much previous research has shown (e.g., Laufer & Goldstein, 2004; Pellicer-Sánchez & Schmitt, 2010; Van Zeeland & Schmitt, 2013), this finding implies that recognition knowledge is easier for L2 learners than recall knowledge. If we interpret recognition knowledge to be closely related to receptive mastery and recall knowledge to better represent productive mastery of an aspect, these results provide evidence for the claim that receptive vocabulary knowledge is higher than productive knowledge in L2 learners (Webb, 2008), confirming the significant difficulty for L2 learners in mastering productive vocabulary knowledge.

Although the pattern of knowledge of each vocabulary component in relation to the others was comparable across both language groups, the results also show that different populations of language learners have slightly distinctive degrees of knowledge of some of these aspects, particularly in recognition. These differences between both groups may be a reflection of their diverse vocabulary size, where Spanish participants generally outperformed Chinese learners. This larger recognition vocabulary size may have led to Spanish learners overall performing better in two of the four recognition components. In addition, it seems intuitive that learners acquiring a L2 that is closer to their L1 would have an advantage in knowledge. This has been previously evidenced for aspects such as the form–meaning link (Elgort, 2013; Rosselli et al., 2012), but not in most other components. In general, the findings suggest that the commonalities between Spanish and English, particularly the cognate status, give the Spanish population a certain advantage in vocabulary recognition knowledge. Nevertheless, the present study also shows that while Chinese and Spanish learners differed significantly in some recognition aspects, they did not appear to differ considerably in their recall vocabulary knowledge. This seems surprising, since it could have been expected that the cognate status of Spanish and English would provide this language population an advantage not only at the recognition level, but mainly in that of recall knowledge. More importantly, the study seems to reflect that achieving recall knowledge is so complex for L2 learners that L2-L1 language similarities are not

sufficient to significantly improve its mastery as compared to learners of other L1s. That is, interestingly, the cognate status of a word is not sufficient to be able to easily extrapolate the recall knowledge from the L1 to the L2. It also seems to indicate that developing recall mastery of vocabulary requires very advanced knowledge (Webb, 2009) and mental connections that may not be simply described by linguistic or metalinguistic knowledge alone.

Overall, the two studies display, for the first time, a clearer portrait of how the knowledge of multiple vocabulary components compare to one another in L2 learning. Recognition of vocabulary overall is considerably better known than vocabulary recall, regardless of the learners' mother tongue. In particular, recognition of the form–meaning link is the aspect that L2 learners master better overall, while knowledge of multiple meanings and derivatives is the most difficult for all of the components, predominantly in the recall mastery.

7.6.2. Interrelationships between word knowledge components

The word knowledge aspects were found to be powerfully intercorrelated with each other in both groups. These intercorrelations were slightly weaker in the Chinese learners, suggesting that for L2 English learners of non-Romance languages establishing the connections between components may be less obvious, probably due to language dissimilarities. Overall, these strong links between word knowledge components in the mental lexicon of L2 learners confirm previous claims of their interrelations (e.g., Gyllstad, 2013; Qian, 2002; Schmitt, 1998; Vermeer, 2001) and indicate that these components must not be attended to as independent entities. From these findings, it can be proposed that when an aspect is mastered, learners can make connections of knowledge that set the basis and facilitate the acquisition of other aspects. This is in line with Henriksen's (1999) unevidenced claim that the development of more complex aspects, that is lexical associations (semantic, orthographic and morphological) between words, supports the improvement of more basic components, such as the form or the form–meaning link. My findings add to this claim, suggesting not only that the acquisition of more difficult aspects help development of simple aspects, but that this also seems to be true the other way around.

These strong intercorrelations also apply to the link between vocabulary size (as operationalised by the VLT) and depth of vocabulary, as represented by multiple components. The relationship between size and depth of vocabulary has been of great interest in the field of vocabulary. Generally conceived as and treated as two different constructs, many previous studies have found high intercorrelations between them (see Qian, 2002; Tannenbaum et al., 2006). My data suggests that vocabulary size is strongly intercorrelated with all depths of vocabulary knowledge aspects, sharing a variance in vocabulary knowledge of between 33% and 80% across both groups. Again, the correlations for the Chinese group were weaker, ranging from .57 to .78 as opposed to the .76 to .90 range in the Spanish group. These results indicate that the greater the vocabulary size of participants, the better the knowledge of vocabulary aspects, although this effect seems stronger for Romance languages. The findings also show that vocabulary size seems to be more strongly related to recall aspects than their recognition counterparts regardless of the learners' L1. This may be an indication that a larger vocabulary size may benefit recall knowledge of words slightly more than their recognition.

In general, the two studies show that the multiple word knowledge components (form–meaning link, derivatives, collocations and multiple meanings, both in recall and recognition) are strongly connected with each other, and also with the learners' vocabulary size, indicating that acquisition of an aspect benefits from the knowledge in any other aspect. Finally, these effects seem in general slightly stronger in learners whose L1 and L2 languages share linguistic features.

7.6.3. General order of acquisition of word knowledge components

Earlier multicomponent research suggested certain trends in the acquisition of word knowledge aspects (e.g., Chen & Truscott, 2010; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Webb, 2007a,b, 2009). Yet, those studies are inconsistent in the aspects they measure and the assessment formats employed, and have typically not been replicated for different L2 populations. Thus, they failed to propose a consistent pattern in the acquisition of word knowledge components. Study 1 in Chapter 4 found preliminary evidence of a consistent vocabulary acquisition order for a sample of L2 leaners. However, in order to claim that there is a general order of

acquisition of English vocabulary in second languages, the pattern needed to be replicated for learners of other languages. The findings from the current replication study verified the same order for a different L2 sample. My results show, for the first time, that EFL learners from diverse linguistic backgrounds follow a systematic and reliable order in the acquisition of various vocabulary knowledge components. This finding is an initial indication of the scale's generalisability, and hypothesises a potential natural acquisition order for vocabulary components in second languages. According to the supported scale, the acquisition of word knowledge components can be expected to adhere to the following sequence in L2 English learners, from early acquired to late acquired aspects (Figure 26).

Figure 26 Generalisable implicational scale of word knowledge components.

Form–Meaning Recognition > Collocation Recognition > Multiple-Meanings Recognition > Derivative Recognition > Collocation Recall > Form–Meaning Recall > Derivative Recall > Multiple-Meanings Recall

From this finding, it follows that recall and recognition of vocabulary knowledge should not be seen as a dichotomy, but rather developing along a continuum. This indicates that in order for learners to acquire recall mastery, and thus produce vocabulary, they need to progress along the scale, mastering recognition first. From the way the word knowledge aspects have been conceptualised and tested in this thesis, recognition can be interpreted as a type of knowledge more closely connected to receptive knowledge and recall as an aspect leading more towards productive mastery. In this sense, my findings serve as evidence for theoretical approaches that understand the relationship between receptive and productive knowledge as a continuum (e.g., Melka, 1997; Read, 2000), where substantial receptive vocabulary knowledge is required before productive mastery is achieved. That is, the recognition knowledge is included and extended in the recall mastery (Henriksen, 1999; Nation, 2013). The present study also shows that learners first master recognition knowledge of aspects such as the form–meaning link, and only later of more complex components, like derivatives.

This generalisable order of acquisition of individual word knowledge components confirms previous research findings that recognition of the form–meaning

link is expected to be acquired the earliest by L2 learners, while recall knowledge of derivatives and multiple meanings would be the last aspects to be mastered (Agustín Llach & Moreno Espinosa, 2014; Barcroft, 2002; Jiang, 2000; Wolter, 2009). My research adds to this earlier research that demonstrating mastery of complex aspects such as recall knowledge of derivatives or multiple meanings would imply having mastered easier aspects such as recognition of the form–meaning link or collocations first. More importantly, this replication study provides the basis for the interpretation of the difficulty of vocabulary components relative to other word knowledge aspects. For example, collocational knowledge has been asserted to be one of the most difficult aspects for L2 learners to acquire (e.g., Chui, 2006; Peters, 2016; Shimamoto, 2000; Webb et al., 2013). However, when compared to less explored aspects, such as derivational and multiple-meanings components, it can be seen that collocational mastery is acquired earlier, in particular when knowledge of very recurrent collocates is required.

The implicational scale shows the same order of acquisition of components for both the Spanish and Chinese groups, despite the former having a larger vocabulary size. The fact that two populations with very distinct L1s and different vocabulary proficiencies follow a stable order of acquisition of vocabulary knowledge components indicates that the scale might reflect a natural order of vocabulary acquisition that remains intact regardless of the participants' proficiency levels or mother tongues. Yet, taking into account the small number of learners that mastered aspects towards the end of the scale, my study is in line with Dóczi and Kormos's (2016) finding that a slowdown tendency in the rate of acquisition of overall vocabulary knowledge is found as the learners' L2 proficiency increases. This trend was particularly noticeable for aspects such as multiple meanings, which is corroborated by my studies. Dóczi and Kormos associated this finding to the power law effect, which predicts that "at higher levels of proficiency vocabulary growth becomes slower" (p. 34). According to this effect, in order to achieve mastery of those complex aspects, more proficient learners would require a considerably greater amount of exposure to the L2, richer, and more varied than lower proficiency students would need to master the easier components.

If learners of different proficiency levels follow the same order of acquisition, it might be that the implicational scale is generalisable not only to other L2 learners, but potentially to L1 learners. Indeed, a study of child native speakers of English found that

depth of vocabulary knowledge develops incrementally over time, starting from the recognition of words, to production and explanation of their meaning, to finally gaining more complex contextual knowledge, including acquisition of multiple meanings of a word and being able to use it in multiple contexts (Tannenbaum et al., 2006). My findings concur with their conclusion that recognition knowledge needs to be acquired before recall mastery is achieved, and that recognition of the form–meaning link is the starting point for overall vocabulary knowledge, whereas learning multiple meanings of words is one of the final stages in this process. Nevertheless, the generalisability of this order to L1 learners is admittedly only a speculation, and thus replications with native speakers are needed to investigate how much of the implicational scale can be extrapolated to L1 vocabulary acquisition.

Finally, my studies present a general learning situation, where various kinds of learners (i.e., different in their proficiency levels and L1s) have learnt different vocabulary from distinct teaching and learning situations. Previous studies using specific learning techniques (deliberate and incidental) have suggested that the knowledge of vocabulary components L2 learners possess is affected by the specific treatment technique employed (Webb, 2005, 2009). Therefore, my scale might admittedly not be applicable to these specific learning conditions. However, I can conclude that when measuring a representative sample of words (different frequencies and cognate status) and not being situated in any particular learning environment, the word knowledge components do seem to be ordered according to how well they are known by at least two different L2 learner populations.

7.6.4. Pedagogical implications

DeCamp (1971) asserted that implicational patterns can be employed to create rules about the acquisition of the linguistic features tested in a specific community. My finding of a general order of acquisition of word knowledge components for L2 learners of English can thus inform of a more systematic and theory-based vocabulary instruction methodology. That is, it can be employed to decide what to prioritise in the curriculum at particular stages of L2 vocabulary learning.

According to the scale, recognition of vocabulary is easier, and thus is expected to be acquired earlier than recall mastery in L2 learners. Practitioners should ensure that

learners are exposed to as much receptive/recognition vocabulary as possible from the beginning of instruction in order to push development along the continuum earlier. The first target of deliberate teaching from that receptive vocabulary input should be mastering recognition of the form–meaning link. Nevertheless, while mastering the form–meaning link aspect, the rich input would also help learners start obtaining information and intuiting other types of recognition knowledge, such as a word's collocations or derivatives. Therefore, when focusing on students' learning of a particular aspect, the target words should be presented in various contexts where the different types of knowledge appear. For example, when focusing on the form–meaning link aspect, it would be useful to present the target word in different sentences and situations so that other kinds of word knowledge come up.

The scale also indicates that a large amount of recognition knowledge is required before a learner can master recall knowledge. Thus, vocabulary instruction needs to get learners to a receptive mastery first. However, it should not stop there. As Melka (1997) suggested, knowing vocabulary is a process by which receptive knowledge moves towards production on a continuum. My research also provides evidence that developing knowledge towards the productive end of the receptive/productive continuum is complex and requires more time. The mean scores in the recognition aspects are quite high for both groups, while they are much lower in the recall aspects. More importantly, the implicational scale shows that the number of people who mastered the recognition aspects overall is much larger than the few people that mastered recall aspects overall (although there are differences for each individual aspect). This suggests that the gap between no knowledge and receptive/recognition knowledge may be smaller than the gap between receptive/recognition knowledge and productive/recall knowledge. If productive knowledge of all components comes later in the learning process and requires considerably more time, thus it would seem a sensible target for pedagogy. This is in line with Schmitt's (2008) and Webb's (2009) conclusion that if productive knowledge is to be developed, it needs to receive a lot of attention and be actively taught through productive learning tasks, rather than just being assumed to follow on automatically from receptive knowledge. For example, Lee (2003) found that her secondary school ESL learners, after being given explicit vocabulary instruction, produced (in compositions) 63.62% of the words they knew receptively, compared to only 13.19% before the instruction. In a delayed composition (23 days later), the

productive percentage had only dropped to 55.46%, showing that the gains were durable (see also Lee & Muncie, 2006). The relatively good retention of productive vocabulary in these studies shows the value of structured productive practice (and encouragement) in helping learners to reach higher levels of productive recall mastery. Thus, in order to push the development from receptive mastery to productive knowledge, teachers need to employ the receptively known aspects in productive learning activities. The pedagogical implication is that language practitioners need to be clear whether they are expecting their learners to have receptive or productive mastery of the target vocabulary, and understand that if productive mastery is expected, it will require having mastered receptive knowledge and considerable time and effort.

My study also emphasises the importance of knowing multiple word knowledge aspects for good vocabulary mastery. The form–meaning link is the aspect that has been shown to receive the most attention in L2 textbooks, and thus, language instruction (Brown, 2011), overlooking other depth of knowledge aspects. My findings suggest that in order to help our students develop good vocabulary knowledge and connections between words, L2 vocabulary instruction should focus on all these types of word knowledge. The proposed order of difficulty can provide the basis for the order in which to gradually introduce the different word knowledge components. This order suggests that the recognition form–meaning link should be the starting point, providing learners with the minimum amount of knowledge to start using the language (e.g., reading simple paragraphs). The rest of the aspects should be introduced incrementally after some knowledge of the recognition form–meaning link has been achieved.

The teaching approach, though, should vary to accommodate to the different characteristics of the aspect to be taught. The learning of form–meaning link and derivatives would benefit from more explicit form-focused instruction, while multiple-meanings and collocational knowledge require a specific contextual situation to be fully understood, and thus would be more amenable to implicit learning. Activities such as matching or fill-in-the-blanks (Laufer & Rozovski-Roitblat, 2011), using word cards or word lists (Elgort, 2011), or translation (Joyce, 2015; Laufer & Girsai, 2008), have been found to be useful for learning the (receptive) form–meaning link of words. Also, the great complexity displayed by derivative knowledge suggests that this aspect should be the centre of explicit instruction. Morphological awareness has been found to lead to vocabulary growth, and thus instruction on morphological awareness can help promote

learners' ability to acquire new words on their own (Kieffer & Lesaux, 2012b; Spencer et al., 2015; Zhang, 2016). Teachers should raise this morphological awareness in students, first by reflecting on the different parts involved in morphologically complex words and the rules of word formation. Once this is acquired, teachers can encourage the use of generative vocabulary activities through the manipulation of suffixes, in which learners create derivatives from words they know following morphological rules. This awareness would allow students to successfully recognise derivational forms of a word when encountered in context. Later, this awareness would enable them to manipulate words to correctly derive their different word classes, which would in turn help consolidate their recognition knowledge of this aspect.

Knowledge of collocations and multiple meanings involves acquisition of contextual information, and thus would be more difficult to teach explicitly, but instead are more easily learnt through implicit instruction and incidental learning. Regarding collocations, although they are not needed for first communication attempts, they facilitate a natural language use and thus communication with other speakers. This knowledge should be enhanced by extensive exposure to words in multiple and varied contexts. As for knowledge of multiple meanings, teachers should go beyond providing a definition or translation of a word, and offer more in-depth, contextual information about the target words (Beck et al., 2013). This way, learners would not only acquire the core meaning of a word, but also develop knowledge of how that word can be used with different senses in different contexts. In order for this multiple-meanings knowledge to advance, learners should be exposed to multiple and varied encounters with words. These repetitions would improve the knowledge and access of the core meaning while developing knowledge of other senses receptively. Guessing from context is considered a useful tool to access knowledge of the multiple meanings of a word, especially if they are based on the core meaning (Webb & Nation, 2017). Therefore, the goal should be to learn the core meaning of words in the L2, so that when encountering that word in a different sense it can be dealt with as something related to the known meaning of that word. Later, generative activities where learners are asked to employ a word in different contexts with its various meanings would pave the way for the development of productive mastery.

Overall, extensive meaning-focused input seems appropriate to help learners master the receptive knowledge of collocations and multiple meanings through reading

and listening, while revising previously known aspects (Henriksen, 1999), like the form–meaning link. To move towards the productive mastery of these components, learners should be exposed to meaning-focused output in the form of conversations or writing stories. These output activities would provide the chance for learners to notice gaps in their productive knowledge of previously and/or partially acquired word knowledge aspects (i.e., spelling recall in the form–meaning link), as well as take risks of using words in ways/aspects they are not completely sure about (Swain, 2005). Since meaning-focused output activities target the productive mastery of aspects that are already known receptively (Nation, 2007), these activities are in line with the implicational scale of vocabulary acquisition order.

Finally, it should be noted that within the systematic pattern there is always some amount of individual learner variation at each stage of acquisition (Ellis, 1989). As such, it is possible that a particular learner may have mastered a complex aspect without having mastered all of the easier ones. Nevertheless, the generalisability shown by this implicational scale, holding across two language populations and scoring criteria, suggests that it is robust enough to be considered useful for pedagogical purposes. In fact, this scale is the best empirically-based description to date of how L2 learners acquire vocabulary. Thus, following this order in instruction would ensure learners receive teaching which closely mirrors what seems a natural acquisition process. However, it is not meant to dictate what the teacher should do in every teaching situation. From this scale, the instruction focus should be contingent on the specific teaching purpose of each class or the learners' proficiency level (see Section 8.6.4, Chapter 8). Overall, this finding shows the pathway learners go through when acquiring vocabulary, and suggests an empirically-based organisation of vocabulary teaching.

7.6.5. Limitations and directions for further research

Replication studies are essential in science to provide validity to previous results. The present study reproduces the findings for the Spanish population and extends them to the Chinese population, offering the implicational scale extra reliability and power, and providing an indication of generalisability of word knowledge

acquisition. Nevertheless, there are certain issues that can be the starting point for future work[21].

First of all, all measures of vocabulary knowledge in this study were given at a single point in time. As such, the implicational scale bases the general order of acquisition of word knowledge components found on the relative difficulty of those items at a particular time. Measuring knowledge of vocabulary components at various stages can inform of whether the expected order of acquisition suggested in this study remains for the same participants over time during the development of vocabulary knowledge. Future research can test overall vocabulary knowledge at two different points in time, and compare the scales found at each of those points, in order to discover whether developmental stages on vocabulary acquisition can be established.

Another limitation is that the findings are replicated for only two populations of language learners. Although this is more than most studies in the vocabulary field have been replicated, it is still not enough to ensure the universality of the acquisition order of word knowledge components. Such a big statement requires further replication studies, with learners of various other mother tongues, in order to certify a universal vocabulary acquisition order. Additionally, it would be interesting to assess whether the order of acquisition followed by these two EFL learner populations also applies to learners in a more naturalistic setting (ESL learners), to younger learners of English (i.e., children) or to native speakers of a language. If the same scale remains in different types of English learners and/or native speakers of English, then the pattern of acquisition of vocabulary knowledge would be confirmed as being universal. These possibilities should be investigated in the future.

This study investigated a learning difficulty order in relation to the word knowledge components in English as a second language. However, the findings may be, in principle, relevant to the acquisition of other second languages. The area of second language vocabulary acquisition would benefit greatly from the application of this study's design on different second languages. If similar results are found for other L2s, it would provide the validity for a general theory of L2 acquisition.

Finally, the general implicational scaling findings are based on the same test battery. As any experimental study, the way knowledge is measured affects the results.

[21] The limitations and suggestions for additional research described for Study 1 in Chapter 4 are also applicable to this study.

I believe the tests employed in these studies are the most appropriate for the purpose and conditions of the studies. Nevertheless, there are multiple ways word knowledge can be assessed. Future research should explore how the order of difficulty of word knowledge components varies when employing different tests or assessing different components. For example, in this study the measures were highly controlled. Although this was needed to enable data collection of multiple word knowledge components concurrently, they do not completely represent productive and receptive skills. Thus, future research could assess these types of knowledge through writing and reading tasks that can better inform of what the learners can do with their vocabulary mastery.

7.7. Conclusion

Most vocabulary research lacks replication. As a result, the findings are accepted as factual on the basis of one test in a specific situation and under certain conditions. Failing to reproduce or refute findings can lead the field to operate under false assumptions. Therefore, it is fundamental that the legitimacy of our findings is checked. The current study undertakes a replication of the implicational scale reported in Study 1 (Chapter 4), in an attempt to investigate whether the order of acquisition of vocabulary components is language dependent or generalisable to other learner populations. Applying implicational scaling and Mokken scaling analyses, the replicability of the order of acquisition of vocabulary components was demonstrated. In other words, the results suggest that the order of difficulty of the word knowledge components tested seems to be the same for the Spanish and Chinese EFL learner populations. This finding provides preliminary indication that the order of acquisition of word knowledge components may be universal. According to this order, EFL learners acquire vocabulary recognition mastery of all word knowledge components before they develop recall mastery. This is in line with previous approaches that suggest receptive knowledge is required before learners can advance towards productive mastery. It also indicates that learners master knowledge of multiple meanings and derivatives in the last place, if they ever achieve such mastery. Theoretically, thus, this replication study provides a general model of how vocabulary knowledge overall is acquired in second languages. Pedagogically, this offers language teachers and practitioners indications for the

development of a systematic vocabulary instruction approach that accommodates to what could be a natural learning order.

Chapter 8

Conceptualising vocabulary knowledge: A replication study

Study 3 in Chapter 6 explored the overall nature of the vocabulary construct, and suggested an empirically-based conceptualisation of vocabulary knowledge. This new conceptualisation indicated that the theoretically-based multidimensionality of vocabulary knowledge was not supported empirically. Rather, structural equation modelling proposed that vocabulary knowledge is best described as a unidimensional construct comprised of different levels of difficulty. The analysis was carried out with a learner sample with Spanish as L1. Hence, although SEM has inferential power, it is questionable whether this conceptualisation would hold for learners of a very different mother tongue. This chapter reports on a replication study that investigates the validity of the unidimensional model of vocabulary knowledge with Chinese EFL learners. Providing this model is supported, it can be considered preliminary evidence for a potential universal conceptualisation of vocabulary knowledge.

8.1. Background of the study

Chapter 6 introduced the issue that vocabulary knowledge is theoretically interpreted as multidimensional despite the lack of supporting empirical evidence for it. It also described the required methodology to investigate the validity of theoretical frameworks (i.e., SEM), and reported a study that explored the multidimensionality of vocabulary knowledge and established an empirically-supported unidimensional conceptualisation of this construct by Spanish EFL learners. Nevertheless, such a finding must be tentative. Although it provides an explanation for the high correlations found between vocabulary components (e.g., Qian, 2002; Vermeer, 2001), the results seem to contest to a certain extent the way vocabulary has been understood for

generations. Thus, it does not seem sufficient to attempt to challenge the multidimensional understanding of vocabulary knowledge based on one learner population. Moreover, Spanish shares a substantial amount of lexis with English. This common vocabulary could have led to higher intercorrelations between the components and, as a consequence, the unidimensional results of vocabulary knowledge. In order to explore whether this conceptualisation of vocabulary knowledge is only valid for EFL learner populations of cognate linguistic backgrounds, Study 3 reported in Chapter 6 needs to be replicated with a learner sample of a non-cognate mother tongue.

The importance of replication studies in the SLA field was discussed in Chapter 7. Still, in studies where the validity and generalisability of a construct is being examined replications are even more essential. Although SEM is considered an approach to validate particular theoretical constructs, proving construct validity for one sample (as I did in Study 3, Chapter 6), more is required in order to generalise that validity to other populations. For this reason, it is vital to conduct replication research to examine the stability of a model across samples and whether the model measures the same construct (In'nami & Koizumi, 2011, p. 270). It is possible that the two studies show that vocabulary knowledge behaves noticeably differently in the mental lexicon of learners from cognate and non-cognate L1 backgrounds, and thus the results would suggest different vocabulary models. Conversely, it might be that vocabulary knowledge behaves generally in a similar manner, suggesting that the unidimensional nature of vocabulary might be universal. The current study presents a replication of Study 3 with a sample of Chinese EFL learners in order to shed light on the generalisability of the unidimensional conceptualisation of vocabulary knowledge and the relationships between the vocabulary components.

In order to avoid repetition, this study moves directly to the methodology section. See Chapter 6 for the background information on multidimensional SEM studies, SEM description and SEM data analysis.

8.2. Purpose and research questions

The present study explores the overall structure of vocabulary knowledge of Chinese learners of English with the aim to establish an empirically-supported conceptualisation of the vocabulary knowledge construct.

Study 5 in the present chapter addressed the following research questions:

1. How strongly do the various word knowledge components contribute to vocabulary knowledge in Chinese EFL learners? What are the similarities and differences between Spanish and Chinese learners regarding the components' contribution to overall vocabulary knowledge?
2. Can predictive relationships be described between certain components?
3. Is there a potential universal conceptualisation of vocabulary knowledge that is empirically valid?
4. If not, how do the vocabulary knowledge frameworks of cognate and non-cognate language learner populations differ?

The next section presents the methodology used to answer these research questions.

8.3. Methodology

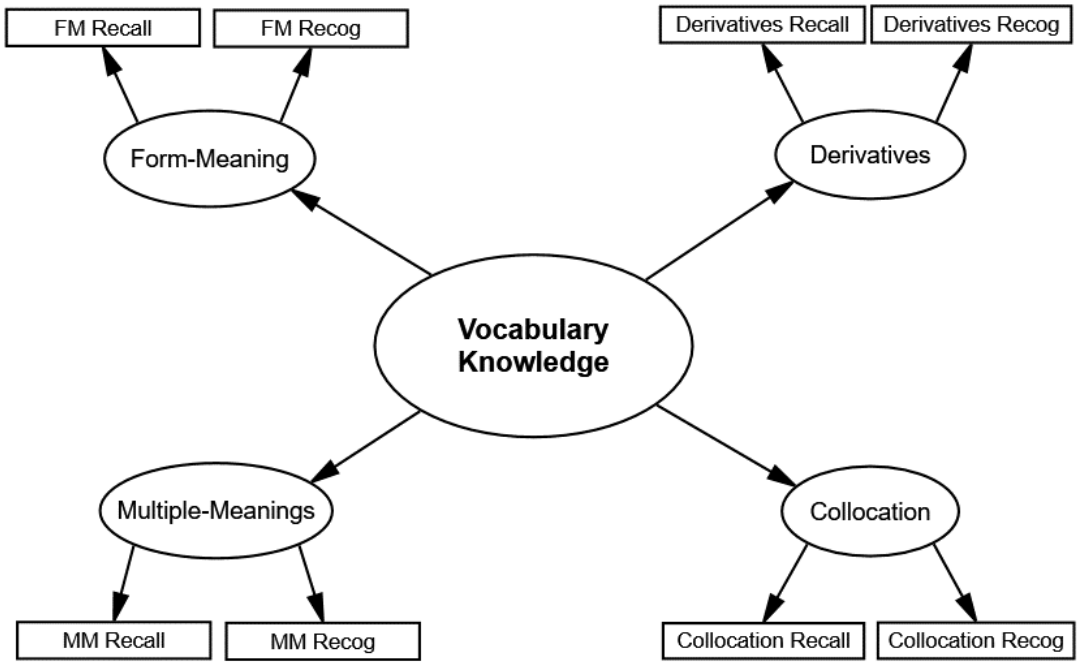
As in Study 3 (Chapter 6), the current study employs the measurement instrument described in Chapter 3. The dataset employed for the analysis was the one obtained for Study 4 (Chapter 7), and thus the participants are those presented in the corresponding section in Chapter 7. In order to recap, this section will briefly describe the hypothesised model submitted for examination and the general process of data analysis. For more details about this, though, refer to Chapter 6.

8.3.1. The hypothesised model

Since SEM is a confirmatory approach, a model of relationships between the vocabulary knowledge components needs to be hypothesised prior analysis. Although Study 3 in Chapter 6 suggested that the best-fitting model of vocabulary knowledge for the Spanish learners was the unidimensional model, there is not enough evidence to simply assume this will also be the case for the Chinese participants. Therefore, the multidimensionality of vocabulary, involving the acquisition of various components, is tested again in Study 5 in the current chapter. Nation's (2013) framework of what it

means to know a word was employed as the basis for building the hypothesised model. He suggests that vocabulary is comprised of multiple word knowledge dimensions, each of which is made up of receptive and productive knowledge. Thus, the hypothesised model in this study, as in Study 3 (Chapter 6), represents vocabulary knowledge as a latent construct comprised of smaller sub-constructs (i.e., form–meaning link, derivatives, multiple meanings and collocations), which, in turn, are represented by their recognition [Recog] and recall [Recall] knowledge. Figure 27 shows the structural equation diagram for this conceptualisation.

Figure 27 Hypothesised model of relationships between word knowledge components.



The application of SEM to test this model will allow us not only to examine the relationships among the word knowledge components and the multidimensionality of vocabulary, but also to establish similarities and differences in the behaviour of vocabulary in the mental lexicon across different learner populations. If this multidimensional model is validated for the Chinese learner population, then it can be claimed that vocabulary knowledge seems to function differently in learners of cognate and non-cognate languages. On the other hand, if this hypothesised model is not supported for this sample and requires respecification into a unidimensional model, the

results from Study 3 (Chapter 6) would be confirmed and an argument for a general conceptualisation of vocabulary knowledge will be made.

8.3.2. Data analysis

For this study, data was collected from an adequate number of participants that allows for SEM analyses (N = 170), which can be considered a reasonable medium-to-large sample (Ding et al., 1995; Kline, 2016). Preliminary analyses were conducted to check for missing values and outliers, as well as to check the normality of the data. Data screening confirmed that there were no outliers or missing data in the variables relevant to the study. The univariate normality of the variables included in the model was tested in Study 4 (Chapter 7), and they showed to be generally non-normally distributed. As in Study 3 reported in Chapter 6, the multivariate normality of the data was assessed by means of Mardia's Multivariate Normality Test (Mardia, 1970) in the R-software with the package *MVN* (version 4.0.2) (Korkmaz et al., 2016). Since the multivariate skewness and kurtosis were $p<.05$, Mardia's test indicated that the data was multivariate non-normal, as expected due to the general nature of social-scientific data (Micceri, 1989; Ramzan et al., 2013). In order to handle the non-normality of the data, robust estimation methods were employed for parameter estimation. As explained in Study 3 (Chapter 6), MLR method was applied for continuous variables, and WLSMV for categorical ones.

The data was analysed using the R-Studio software (R Core Team, 2017), version 3.4.1. Structural equation models were fitted using the *lavaan* package, version 0.5-23.1097 (Rosseel, 2012). The acceptability of the models was judged following the same guidelines as in Study 3 (Chapter 6). A good model fit was indicated by a nonsignificant chi-square (χ^2), chi-square/*df* ratio (χ^2/df) of between 1 and 3, comparative fit index (CFI) of >.95, standardised root mean square residual of (SRMS) <.08, and root mean square error of approximation (RMSEA) of <.05 for good fit and <.08 for acceptable fit.

As in Study 3, the analysis followed three steps. Firstly, the fit of each latent word knowledge component and its indicators was assessed separately by means of CFA, in order to test whether the measurement instruments of the individual word knowledge components indeed represented the aspects I intended to measure. Secondly,

a pooled CFA was applied to the latent constructs and their two indicators as represented in the hypothesised model, in order to test the validity of each latent variable. Lastly, SEM was used to test the multidimensionality of vocabulary knowledge by testing the hypothesised model in Figure 27.

Finally, in order to compare the results between the Chinese sample in this study and the Spanish sample in Chapter 6, Multigroup CFA was applied using the software AMOS 24 (Arbuckle, 2016) in SPSS 24 (IBM Corp., 2016).

8.4. Results and interpretation

8.4.1. Descriptive statistics and correlations

Table 71 summarises the mean percentage scores by the Chinese participants in each of the vocabulary components, as well as the intercorrelations between them. It can be seen that the learners knew at least half of the items in each test on average, being the recognition scores always higher than the recall scores for every individual component. The gap between the minimum and maximum scores indicates that learners had a range of proficiency levels. Spearman's correlation analysis shows that the intercorrelations between word knowledge components ranges between .44 and 72, showing a high degree of interconnection between the components. This supports my previous findings that the word knowledge components are highly interrelated.

Table 71 Descriptive statistics (%) and intercorrelations between components (Chinese group).

	FM Recall	FM Recog	Deriv Recall	Deriv Recog	MM Recall	MM Recog	Collo Recall
FM Recall							
FM Recog	.440**						
Deriv Recall	.483**	.579**					
Deriv Recog	.439**	.551**	.711**				
MM Recall	.611**	.570**	.721**	.656**			
MM Recog	.466**	.620**	.602**	.592**	.722**		
Collo Recall	.454**	.492**	.521**	.533**	.588**	.550**	
Collo Recog	.433**	.485**	.561**	.493**	.603**	.497**	.701**
Min.	10.00	50.00	28.75	27.50	25	28.33	20.00
Max.	85.00	100	83.75	88.75	80.00	91.67	100
M	53.15	79.38	51.57	65.82	50.29	64.90	57.21
SD	11.99	9.53	10.26	10.74	10.00	11.62	15.38

Spearman: **$p < .01$, $N = 170$

FM Recall = Form–meaning Recall; FM Recog = Form–meaning Recognition; Deriv Recall = Derivative Recall; Deriv Recog = Derivative Recognition; MM Recall = Multiple-Meanings Recall; MM Recog = Multiple-Meanings Recognition; Collo Recall = Collocation Recall; Collo Recog = Collocation Recognition.

8.4.2. Structural equation modelling analyses

8.4.2.1. Validation of measurement instruments with individual words

As it is recommended in SEM analyses, a series of CFA were first conducted in order to validate each word knowledge component independently before modelling their relationships in the hypothesised SEM model. Eight independent CFA were tested, one for each word knowledge component. In these models, the results of the 20 target words used to measure each of the recall and recognition aspects were loaded as indicators into their corresponding word knowledge component, which were described as a latent variable (see Section 6.5.2.1 in Chapter 6 for a detail description and illustration of this process). The scores for the individual target words in each test were categorical, and therefore the estimator WLSMV (robust version of the WSL estimator,

which is more appropriate for categorical, non-normally distributed data) was employed.

Due to space constraints, only the model fit indices and reliability of these CFA models are reported. Table 72 shows that the different measures of word knowledge aspects obtained good general model fit values and reliability. This indicates that the target words included in the tests are measuring the concept they were designed to measure, and confirm that the different word knowledge components' tests are valid psychometric constructs.

Table 72 Fit indices and reliability for the CFA models with 20 target words as indicators (Chinese group).

	Model fit indices						
	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	Reliability
Acceptable fit			>.05	1-3	>.95	<.05 / <.08	α >.7
Form–Meaning Recall model	156.299	136	.112	1.15	.94	.03 (.000-.049) *p* = .959	α = .72
Form–Meaning Recognition model	176.010	157	.142	1.12	.96	.03 (.000-.046) *p* = .981	α = .86
Derivative Recall model	224.198	167	.002	1.34	.95	.05 (.028-.060) *p* = .697	α = .85
Derivative Recognition model	219.997	164	.002	1.34	.95	.05 (.028-.060) *p* = .700	α = .84
Multiple-Meanings Recall model	210.372	163	.007	1.29	.95	.04 (.023-.057) *p* = .808	α = .86
Multiple-Meanings Recognition model	209.492	168	.016	1.25	.95	.04 (.017-.054) *p* = .885	α = .85
Collocation Recall model	168.823	160	.301	1.06	.96	.02 (.000-.040) *p* = .996	α = .81
Collocation Recognition model	166.834	160	.430	1.04	.98	.01 (.000-.037) *p* = .998	α = .76

The table also shows that the Form–meaning Recall and Collocation Recognition models have the lowest reliability coefficient, although still adequate. This indicates that these tests obtained the more variation in responses, that is, were answered

by participants in a less consistent manner. It can also be seen that four of the models yielded a significant chi-square *p* value (Derivative Recall and Recognition and Multiple-Meanings Recall and Recognition). As seen in Study 3 (Chapter 6), it is possible to have a valid model even with a significant chi-square if the other fit indices are adequate (Kenny et al., 2014). For this reason, some researchers advise reporting the chi-square information, but ignoring it when deciding on the goodness-of-fit of the model (Awang, 2012). The rest of the fit indices suggest a good model fit, and thus these models were considered valid and reliable.

The different number of degrees of freedom for each model reflects correlations between the measurement errors for some words. The evaluation of the models suggested that these correlations between target words, or deletion of some items with low factor loadings, were desirable for a good model fit. Removing worse-performing target words is not an acceptable solution for the purpose of this study for two main reasons. First of all, difficult and easier words were included in the test to obtain a better representation of vocabulary, and to attend to the different proficiencies of participants. It was anticipated, and necessary in this study, that the easy words were known by most people, and the difficult words unknown by most. Thus, deleting these items would lead to a loss of valuable information (refer to Chapter 6, Section 6.5.2.1 for a detailed explanation of these reasons). Second, since this is a replication study, the same conditions as in Study 3 (Chapter 6) need to be reproduced. Thus, for these two reasons, item error correlation was the most suitable solution. After this modification, it was observed that the error correlations occurred mainly within easy or complex items. For example, in the form–meaning recall test, *redeem* and *grate*, *development* and *indent*, or *terminal* and *indent* correlated highly with each other, as well as *fresh* and *mean*. Or in derivative recall test, *draught* and *grate*, and *grate* and *indent* were found to correlate. Thus, the words that benefited from error correlation were those particularly easy or complex words.

Overall, the measurement instruments used to represent the different word knowledge components showed to be reliably representing the general construct they were designed to measure.

8.4.2.2. Validation of latent variables for the hypothesised model

After individually testing the validity of the different word knowledge aspects, I validated the latent variables as displayed in the hypothesised model. In this model, the recall and recognition aspects of each word knowledge component were specified as two indicators, each of which were comprised of the composite total test score of the 20 target words employed to assess them (see Chapter 6, Footnote 17 for an explanation of this process). Thus, each word knowledge component is represented by two indicators, reflecting their recall and recognition knowledge. In order to inspect the connection of the recognition and recall indicators to their respective factors, I first evaluated the fit of the measurement models by conducting a pooled-CFA of all the factors included in the hypothesised model (see Figure 28). This analysis was also employed to assess the unidimensionality, validity and reliability of each latent construct before assessing the hypothesised model. The MLR estimator was employed, due to the continuous and non-normally distributed nature of the variables in the hypothesised model. Figure 28 shows that each measurement model was satisfactorily established, with all indicators being significant. The relationships among latent variables were also all significant.

Figure 28 Pooled-CFA of all factors in hypothesised model (Chinese group).

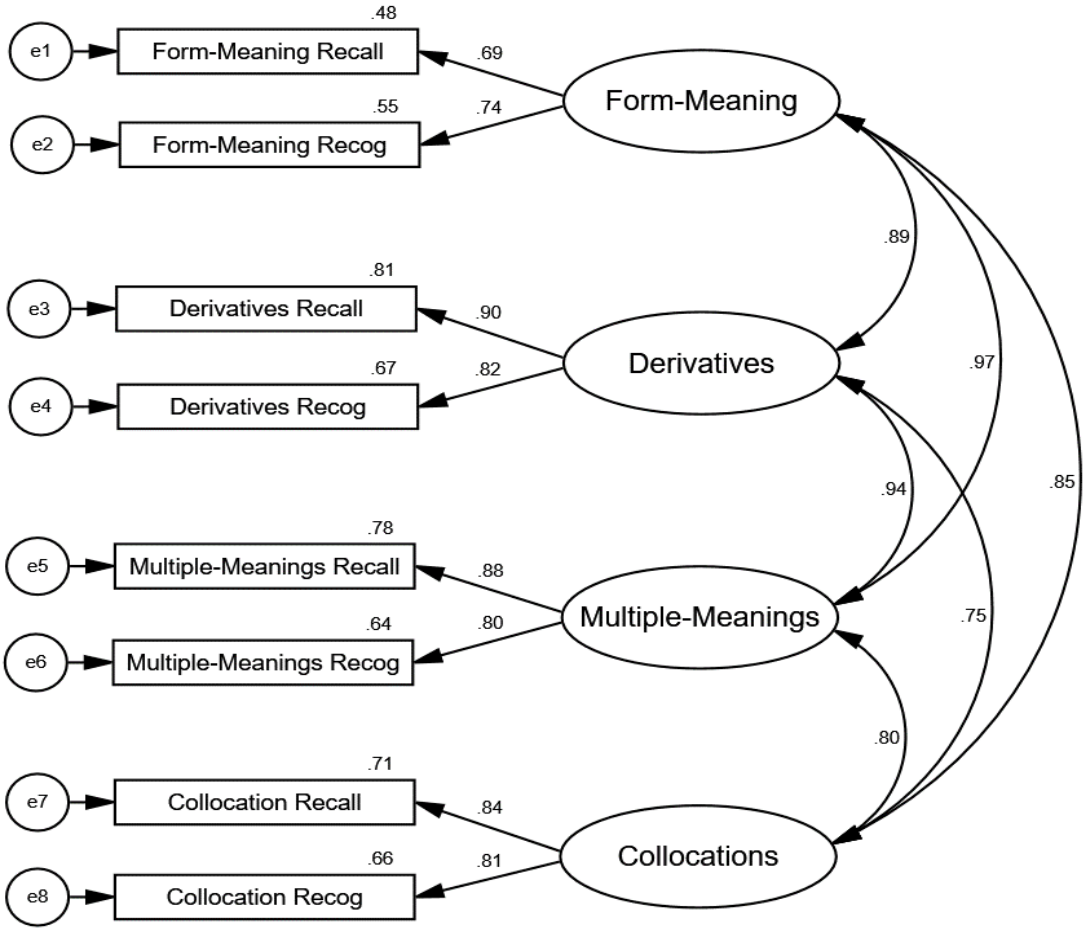


Table 73 shows the fit indices of the Pooled-CFA for the Chinese population. The model's χ^2 was 13.644 (df = 14; p = .48), and the ratio of χ^2/df was .98, which indicates generally acceptable fit. The RMSEA of .00 and the CFI of 1.00 suggest a near-perfect model fit. Finally, the SRMR of .015 was very close to 0, and thus confirms the good fit.

Table 73 Pooled-CFA fit indices (Chinese group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			p>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	13.644	14	.478	.98	1.00	.00 (.000-.072) p = .81	.015

The fit indices of the pooled-CFA reached the threshold to consider the model valid, and thus construct validity was achieved. There is only one value that is not as high as might be hoped: the ratio between chi-square and degrees of freedom, which is just approaching the minimum value of 1. Nevertheless, since it is so close to the acceptable threshold and the rest of the fit indices suggest a very good model fit, there seems to be enough evidence to consider the model valid. Unidimensionality was achieved by each factor (factor loadings >.40). Convergent validity was also attained by every factor, as represented by the average variance extracted (AVE) (Form–meaning AVE = .52, Derivatives AVE = .74, Multiple-Meanings AVE = .71, Collocation AVE = .69). Finally, the overall model was found to be reliable (Cronbach's α = .88), as well as the measurement models individually (Form–meaning Cronbach's α = .67, Composite Reliability (CR) = .68, Derivatives Cronbach's α = .85, CR = .85, Multiple-Meanings Cronbach's α = .82, CR = .83, Collocations Cronbach's α = .80, CR = .81) (see Table 74).

Table 74 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for Pooled-CFA model (Chinese group).

Factors	**AVE**	**CR**	**α**
Acceptable value	>.5	>.6	>.7
Form–Meaning link	.52	.68	.67
Derivatives	.74	.85	.85
Multiple-Meanings	.71	.83	.82
Collocation	.69	.81	.80
Overall model			.88

Table 75 displays the factor loadings, squared multiple correlations, *z* scores and significance for the pooled-CFA model. Since all paths were significant, convergent validity was confirmed. The squared multiple correlation indicates that the reliability of the indicators represented by composite scores is good (>.70), and by extension, the factors are reliable constructs.

Table 75 *Z* scores and significance for Pooled-CFA (Chinese group).

Paths	**Factor Loadings R**	**R^2**	***Z***
Form–Meaning → Form–Meaning Recog	.74	.55	8.902***
Form–Meaning → Form–Meaning Recall	.69	.48	8.432***
Derivatives → Derivatives Recog	.82	.67	14.137***
Derivatives → Derivatives Recall	.90	.81	21.395***
Multiple-Meanings → MM Recog	.80	.64	11.751***
Multiple-Meanings → MM Recall	.88	.78	14.083***
Collocation → Collocation Recog	.82	.66	12.327***
Collocation → Collocation Recall	.84	.71	12.695***
	Correlations (R)		***Z***
Form–Meaning ←→ Derivatives	.89		13.332***
Form–Meaning ←→ Multiple-Meanings	.97		14.632***
Form–Meaning ←→ Collocations	.85		14.541***
Derivatives ←→ Multiple-Meanings	.94		29.970***
Derivatives ←→ Collocations	.75		14.441***
Multiple-Meanings ←→ Collocations	.80		16.341***

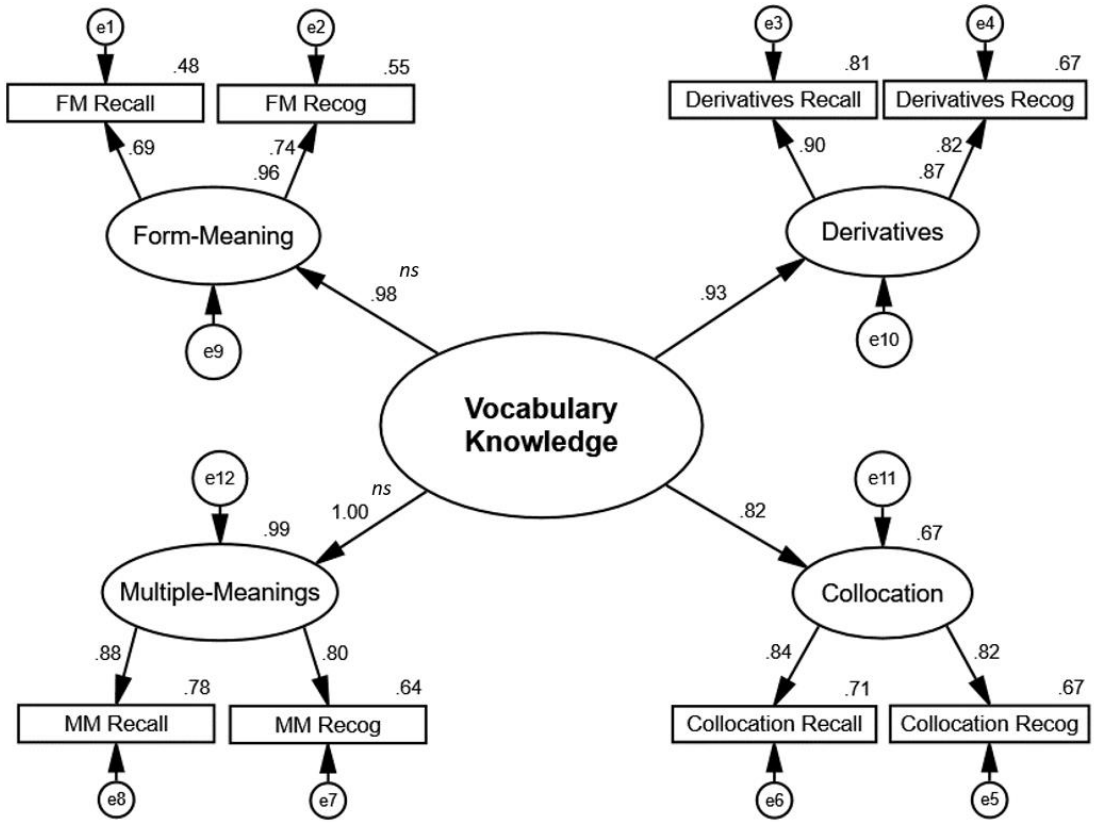
*** p <.001.

The Pooled-CFA results for the Chinese group attest that the recall and recognition indicators reliably represent the construct of each word knowledge component, and thus these components are suitable to be included in the hypothesised model analysis. Consistent with the vocabulary theory, in this model it was expected that the four vocabulary dimensions would be intercorrelated with each other. The results show that this is indeed the case. However, the very high intercorrelations among the different factors (.75-.97) suggest that the constructs are representing very similar concepts. According to Kline (2016), this might reflect redundancy, and thus, lack of discriminant validity between the components. This will be discussed in the hypothesised model evaluation.

8.4.2.3. Hypothesised model

Once the measurement models were validated by means of CFA, the hypothesised model was fitted to SEM employing the MLR estimator. This model, based on Nation's (2013) framework of vocabulary knowledge, illustrates a higher-order structural model in which overall vocabulary knowledge was symbolised by four word knowledge factors represented by their recall and recognition masteries. Figure 29 shows the strength of the relationships between the word knowledge components and vocabulary knowledge as described in this model.

Figure 29 Model 1: Hierarchical model of Vocabulary Knowledge (Chinese group).



Note: All values $p<.01$, except *ns*=nonsignificant ($p=.45$ and .76).

Table 76 shows the fit indices for the hypothesised model. The χ^2 for the model was 16.117 ($df = 16$; $p = .49$), and the ratio of χ^2/df was 1.01, which indicates very good fit. RMSEA of .00 and CFI of 1.00 not only met the standards for acceptable fit, but also indicate almost perfect fit (i.e., near 0 and 1.0 respectively). Finally, SRMR of .019

was close to 0, showing very good fit. Since all the model's fitness indices met the acceptability requirements, construct validity was achieved.

Table 76 Model 1: Fit indices (Chinese group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	16.117	16	.49	1.01	1.00	.00 (.000-.068)	.019
						p = .83	

All factor loadings were way above .40, indicating that unidimensionality was achieved. Convergent validity was also attained, as represented by AVE of the overall Vocabulary Knowledge construct (AVE = .87) and the four sub-constructs (Form–meaning AVE = .52, Derivatives AVE = .74, Multiple-Meanings AVE = .71, Collocation AVE = .69). The overall model was found to be reliable (Cronbach's α = .88, CR = .97), as well as the measurement models individually (Form–meaning Cronbach's α = .67, CR = .68, Derivatives Cronbach's α = .85, CR = .85, Multiple-Meanings Cronbach's α = .82, CR = .83, Collocations Cronbach's α = .80, CR = .82).

Table 77 Average Variance Extracted (AVE), Composite Reliability (CR) and Cronbach's alpha for hypothesised model (Chinese group).

Factors	AVE	CR	α
Acceptable value	>.5	>.6	>.7
Form–Meaning link	.52	.68	.67
Derivatives	.74	.85	.85
Multiple-Meanings	.71	.83	.82
Collocation	.69	.82	.80
Overall model	.87	.97	.88

Table 78 Relative contribution of the different aspects to vocabulary knowledge, *Z* scores and significance (Chinese group).

Paths	**Factor Loadings R**	**R^2**	***Z***
Form–Meaning → Form–Meaning Recall	.69	.48	.752 (*p* =.452)
Form–Meaning → Form–Meaning Recog	.74	.55	.749 (*p* =.454)
Derivatives → Derivatives Recall	.90	.81	5.614***
Derivatives → Derivatives Recog	.82	.67	5.543***
Multiple-Meanings → MM Recall	.88	.77	.310 (*p* =.756)
Multiple-Meanings → MM Recog	.80	.64	.312 (*p* =.755)
Collocation → Collocation Recall	.84	.71	8.419***
Collocation → Collocation Recog	.82	.67	9.294***
	Standardised Regression weights		***Z***
Vocabulary Knowledge → Form–Meaning	.98		.751 (*p* =.453)
Vocabulary Knowledge → Derivatives	.93		4.895***
Vocabulary Knowledge → Multiple-Meanings	1.00		.309 (*p* =.757)
Vocabulary Knowledge → Collocations	.82		6.877***

*** p <.001.

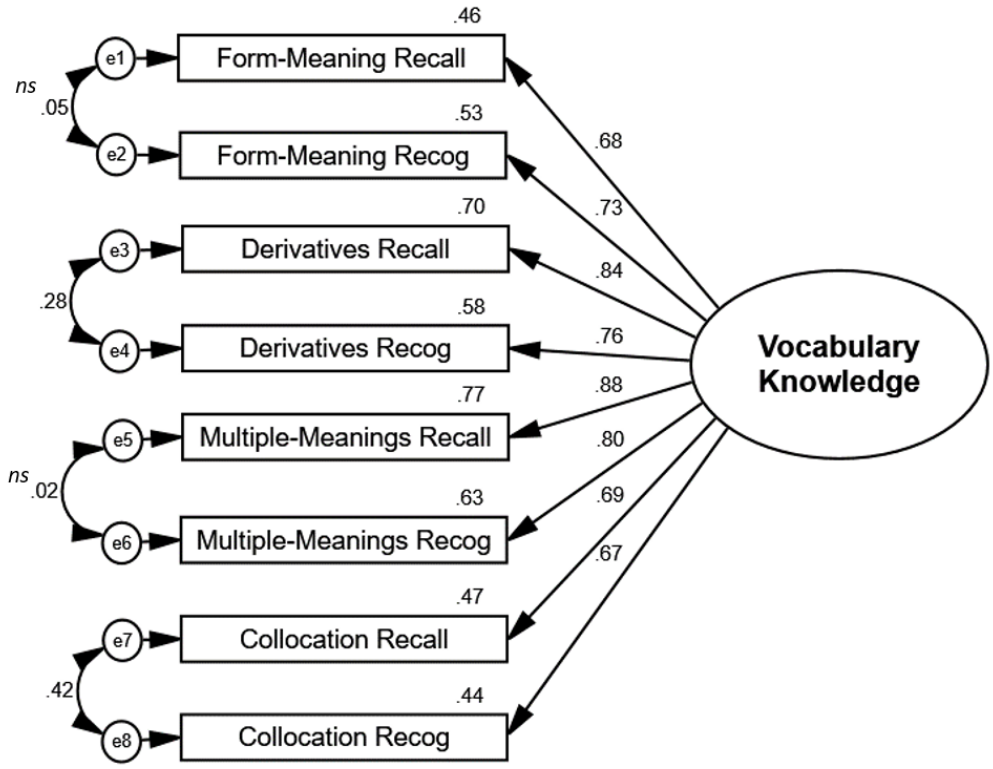
The hypothesised model, or Model 1, shows a good overall fit with the data. However, inspection of the internal structure of the model suggest that the model requires respecification. The regression coefficients between Vocabulary Knowledge and the four word knowledge components show very high values (β = .82-1.00). This suggests lack of discriminant validity between the word knowledge components and affects the significance of some paths (see Table 78). That is, as in Study 3 (Chapter 6), these extremely high β coefficients indicate that the components are very strongly intercorrelated, and thus, empirically indistinguishable from each other. As a consequence, it cannot be claimed that the word knowledge components behave as different sub-constructs of vocabulary knowledge. Examination of the proportion of

variance in each factor explained by vocabulary knowledge (i.e., number on top of each oval) shows that it ranges between 67% and 99%, also suggesting redundancy between the word knowledge components. The collocations dimension experienced the lower regression coefficient (β = .82), which indicates that it is the most discernible aspect. Nevertheless, the regression is still very high. Therefore, despite the good fit indices shown for Model 1, the high regressions between the components and Vocabulary Knowledge suggest that they are redundant and comprise a single construct, and thus the interpretation of the coefficients for this model is meaningless. To sum up, this model is not well supported by the data, and cannot be considered a good representation of vocabulary. Respecification of the hypothesised model is required, removing the redundant variables (Hooper et al., 2008).

8.4.2.4. Revised model

Evaluation of the hypothesised model suggested the need for respecification in the same way as in Study 3 in Chapter 6. Model 1 divided the vocabulary knowledge construct into multiple dimensions, and was not well supported by the data due to lack of discriminant validity between the modelled dimensions. Thus, the model was reconceptualised as in Study 3, building the recognition and recall aspects as individual direct indicators of the general vocabulary knowledge construct, hence eliminating the redundant latent dimensions. In addition, although the recall and recognition aspects behave differently from each other, they are still interrelated, as Table 71 and previous research has found (e.g., Laufer & Paribakht, 1998). Therefore, I specified that the recall and recognition aspects of the same component (e.g., Collocation Recognition and Recall) would intercorrelate. CFA was used to test the fit of the data to the revised one-factor model. This model is illustrated and analysed in Figure 30.

Figure 30 Model 2: Recognition and recall knowledge as independent aspects of vocabulary knowledge (Chinese group).



Note: All values $p<.001$, except *ns*=nonsignificant ($p=.71$ and .88).

Table 79 Model 2: Fit indices (Chinese group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			$p>.05$	1-3	>.95	<.05 / <.08	<.08
Model fit	16.117	16	.49	1.01	1.00	.00 (.000-.068) $p = .84$	.019

The model's χ^2 (16.117; $df = 16$; $p = .49$) and the ratio of χ^2/df (1.01) show a good fit. RMSEA of .00 shows a perfect fit, as well as the CFI of 1.00. SRMR of .019 was very close to 0. All the model fit indices reached and exceeded the commonly accepted thresholds, supporting the suitability of the revised model, and thus, the validity of the construct.

The parsimony fit indices (AIC and BIC), which were designed to compare models and select the best-fitting one, were employed to compare the goodness-of-fit of the hypothesised and revised models. These indices do not seem to be different, suggesting both models are equally good regarding their fit indices. However, although the multidimensional model indicates a good underlying description of vocabulary, the lack of discriminant validity makes it a worse model of vocabulary knowledge.

Table 80 Model comparison for parsimony (Chinese group).

Model	AIC	BIC	Sample-sized adjusted BIC
Model 1	6,994.50	7,082.30	6,993.64
Model 2	6,994.50	7,082.30	6,993.64

As in the previous models, the assumptions of unidimensionality, convergent validity and reliability were tested. All factor-loadings are positive and above the appropriate level (.40), which suggests the model achieved unidimensionality. Besides, all loadings are statistically significant and AVE exceeds .50 (AVE = .58), showing that convergent validity is achieved by this model (Awang, 2012). The reliability and internal consistency of the model was checked by means of Cronbach's α and Composite Reliability (CR). For Model 2, both Cronbach's α and CR showed to be very good (α =.88, CR = .92 (CR>.60 is considered good (DeVellis, 2012)). This suggests Model 2, in which the vocabulary knowledge construct explained by the word knowledge aspects as indicators is a valid and reliable construct. In sum, the relationships between components were satisfactorily verified and significant, demonstrating the validity and reliability of the model.

Table 81 shows the standardised factor loading and the square multiple correlations to see the variance explained by each aspect.

Table 81 *Z* scores and significance of word knowledge aspects as indicators of vocabulary knowledge (Model 2) (Chinese group).

Paths	**Factor Loading R**	**R^2**	***Z***
Vocabulary Knowledge → FM Recall	.68	.46	9.420***
Vocabulary Knowledge → FM Recog	.73	.53	10.518***
Vocabulary Knowledge → Deriv Recall	.84	.71	12.476***
Vocabulary Knowledge → Deriv Recog	.76	.58	10.891***
Vocabulary Knowledge → MM Recall	.88	.77	13.902***
Vocabulary Knowledge → MM Recog	.80	.64	11.142***
Vocabulary Knowledge → Collo Recall	.69	.48	10.488***
Vocabulary Knowledge → Collo Recog	.67	.45	9.469***
	Error Correlations		***Z***
FM Recall ←→ FM Recog	.05		.376 *p*=.71
Derivatives Recall ←→ Derivatives Recog	.28		4.564**
MM Recall ←→ MM Recog	.02		.156 *p*=.88
Collocation Recall ←→ Collocation Recog	.42		4.962***

** $p < .01$, *** $p < .001$.

The factor loadings indicate that all the individual word knowledge aspects load highly (>.40) on the Vocabulary Knowledge construct. These *R* values range from .67 (R^2 = .45) to .88 (R^2 = .77), thus explaining variance in vocabulary knowledge between 44% and 77%. The lowest factor loading is found for Collocation Recognition knowledge (*R* = .67), which is still a very high correlation. This indicates that when we take all these vocabulary knowledge aspects into account, recognition knowledge of collocations explains 44% of overall vocabulary knowledge. The highest correlation is between vocabulary knowledge and Multiple-Meanings recall (*R* = .88), which suggest that the greatest amount of variance in vocabulary knowledge is explained by recall knowledge of multiple meanings (R^2 = .77). This is in line with Tannenbaum et al. (2006), who found that recall knowledge of multiple meanings was the strongest contributor to depth of vocabulary knowledge (.69) when compared to other aspects. The next highest factor loading is found by Derivative Recall (*R* = .84), explaining 71% of the variance in Vocabulary Knowledge when controlling for other word knowledge aspects. It is apparent that the recall aspects tend to contribute to vocabulary knowledge

a bit more strongly than their recognition pairs, except for the Form–Meaning aspect. Yet, in general the factor loadings are high for all the word knowledge aspects, indicating that they all contribute to vocabulary knowledge almost at the same level and thus are equally important in explaining this construct. Thus, the model and its high factor loadings indicate that the word knowledge aspects are very highly intercorrelated with each other and belong to one common construct, Vocabulary Knowledge.

Finally, as in Study 3 (Chapter 6), the correlated errors between the recognition and recall aspects were significant only in the case of Derivatives and Collocations (r =.28, p = .004, and r = .42, p = .000, respectively, see Table 81). This finding indicates that their indicators are measuring something in common not explained by the model. That is, there is overlapping content between each pair. In the case of derivatives knowledge, the error correlation might indicate the need of morphology and grammar awareness to master derivatives (Kieffer & Lesaux, 2012b), while collocational knowledge would involve mastering information about the contextual use of words and acquiring native-like intuitions of it (Siyanova-Chanturia & Spina, 2015). In turn, the lack of significant error correlation between the Form–Meaning and Multiple-Meanings recognition and recall indicators suggests that their variance is largely explained in the overall Vocabulary Knowledge model.

To conclude, Model 2 proves to be the best representation of vocabulary knowledge for the data in this study, suggesting that vocabulary is better seen as a unidimensional model. All word knowledge aspects make a large contribution to the explanation of the Vocabulary Knowledge construct, which demonstrates they are all essential components of knowing vocabulary.

8.4.2.5. Alternative models

Due to the replication nature of the present study, all the exploratory models tested in Study 3 (Chapter 6) were also checked in the present study. Nevertheless, the lack of discriminant validity found in the hypothesised model also for the Chinese group anticipated that the predictive models would also suffer from redundancy.

First, following on the notion that the learners' vocabulary size (generally operationalised as the form–meaning link) can be an indication of the learners' depth of vocabulary knowledge (e.g., Schmitt, 2014), I built a model in which the form–meaning

construct predicted knowledge of derivatives, multiple meanings and collocations (see Figure 31).

Figure 31 Direct effect of form–meaning link on depth of knowledge aspects (Chinese group).

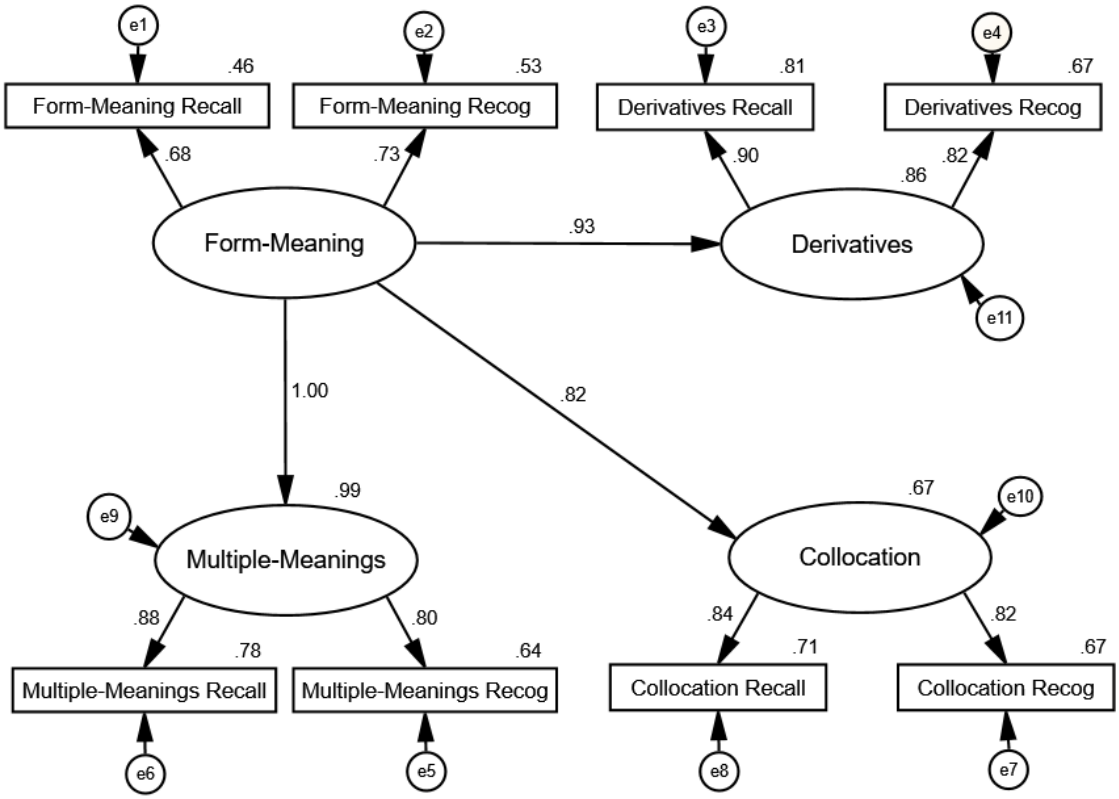


Table 82 Model fit for Alternative Model 1 (Figure 31) (Chinese group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	14.098	14	.48	1.01	1.00	.00 (.000-.072) *p* = .81	.015

Table 82 shows that this Alternative Model 1 has a good fit. However, the lack of discrimination between the different components as independent dimensions (reflected in the regression scores (β = .82-1.00)) indicates that this predictive representation of vocabulary is not supported by my data. If two aspects cannot be

considered essentially distinct from each other, the predictive power of one over the other is meaningless.

Second, in order to test whether the recognition of overall vocabulary knowledge influenced the recall knowledge, I fitted a model in which the latent recognition knowledge, comprised of all the recognition tests, predicted recall knowledge, formed by the four recall tests (see Figure 32).

Figure 32 Direct effect of recognition on recall masteries (Chinese group).

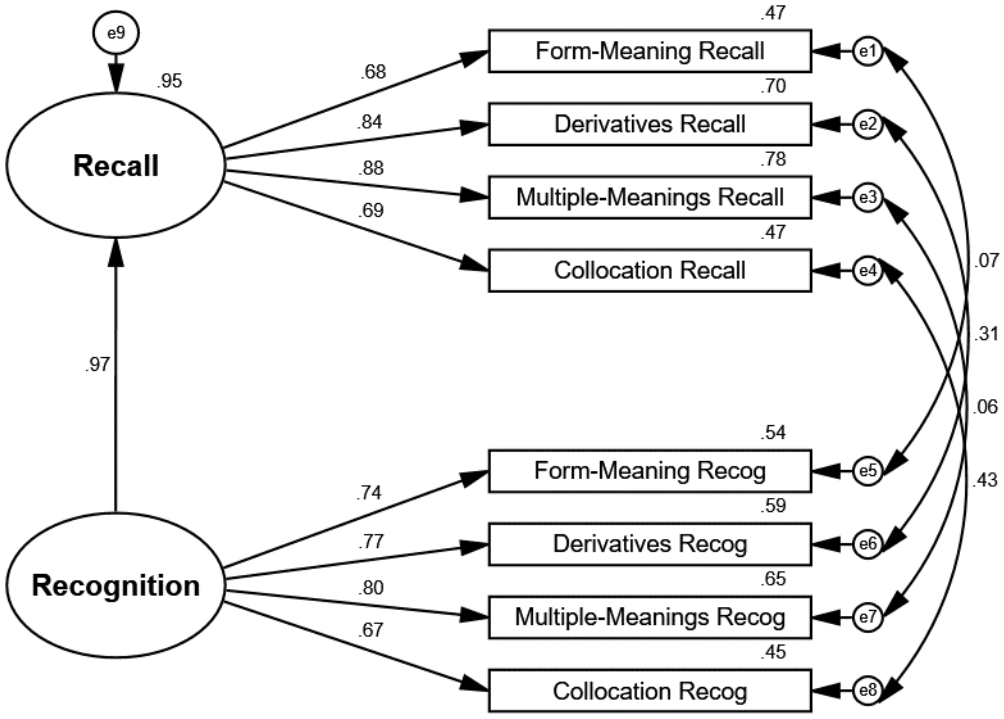


Table 83 Model fit for Alternative Model 2 (Figure 32) (Chinese group).

	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA (90% CI)	SRMR
Acceptable fit			*p*>.05	1-3	>.95	<.05 / <.08	<.08
Model fit	13.833	15	.54	.92	1	.00 (.000-.068) *p* = .85	.019

Yet again, despite the good fit displayed in Table 83, the regression coefficient between the recognition and recall latent variables (β =.97) suggest that they are not separate dimensions. Hence, this causal effect cannot be supported by the data.

These alternative models exemplify interesting relationships between the vocabulary knowledge components. However, the lack of discriminant validity found between these components in the hypothesised model, and confirmed by the high regression values in the above models, prevent the word knowledge components from any predictive power. Therefore, causal relationships cannot be tested between the word knowledge components, since they cannot be considered independent factors. As a consequence, any further interpretation of the alternative models in this section would be meaningless.

8.5. Comparison of results between the Spanish and Chinese groups

In this section, the SEM models from the Spanish and Chinese learners will be compared in order to examine the differences and similarities in their behaviour regarding vocabulary knowledge. The differences between the two language groups regarding correlations and mean scores were already discussed in Chapter 7 (Section 7.5).

8.5.1. Measurement instruments with individual words

The validity and reliability of the individual word knowledge measures was demonstrated for both the Spanish and the Chinese groups, indicating that each individual test is measuring what they were supposed to measure regardless of the L2 learner group taking the tests. Yet, a close comparison of the results showed that the reliability of each component for the Chinese learners, although acceptable, was always slightly lower than for the Spanish learners. The Cronbach alpha for the individual recall and recognition measurement models in the Spanish group ranged between .90 and .96, while in the Chinese group it was between .72 and .86. This indicates that, while the individual measurement instruments are reliably assessing the construct pursued for both groups, the Spanish group as a whole experienced somewhat less variability in the participants' answers to the test. It has been acknowledged that individual tests often

differ in how they operate in different groups of participants (Byrne, 2016). This result shows that there seem to be some differences in the performance of these groups on the word knowledge tests.

However, common patterns across both populations in the behaviour of the tests were found to be more typical. In both groups there were some words that did not seem to have a very strong contribution in explaining the specific concept measured by the tests. This is reflecting the situation that some words are generally easier or more difficult than others, and thus known or unknown by most participants. Error correlations between target items were required for both groups in these cases, representing overlapping content between the words and the fact that in both groups easy words were known by most learners and difficult words were generally unknown. Finally, for both language groups the derivatives and multiple-meanings models were found to be usually the most consistent and reliable tests, whilst the form–meaning and collocation models overall involved more discrepancies. This was reflected in the alpha values of the models, but also in the modifications they required (e.g., error correlations). This indicates a common trend in how the different word knowledge tests performed in the two groups.

Overall, the individual recall and recognition measurement instruments showed a good general representation of the word knowledge components for both groups. Also, although with some slight variations, this comparison demonstrated a common tendency across both group in their behaviour to the different vocabulary knowledge tests.

8.5.2. Latent variables for the hypothesised model

A pooled-CFA analysis was employed to test the validity of the latent variables included in the hypothesised model and their indicators. This analysis demonstrated that the word knowledge aspects as latent variables represented by their recognition and recall indicators had a good fit for both language groups. Thus, they are appropriate representations of the concepts I wanted to measure. In this case, the model fit was slightly better for the Chinese population than for the Spanish one (i.e., non-significant χ^2 p value, and better RMSEA, CFI and SRMR). This might have been a reflection of the larger sample size in the Chinese group (i.e., large samples are preferred in SEM

analyses, (Kline, 2016)). Yet, in both groups, the goodness-of-fit was more than acceptable.

When looking at the internal structure of the pooled-CFA models, slightly lower correlations between the latent variables are seen for the Chinese group (range .75-.97) than for the Spanish one (range .89-.96). Correlations between derivatives and multiple meanings were *r* =.94 in both groups, and between form–meaning and multiple meanings they were almost identical (*r* =. 97 in the Chinese sample and *r* =.96 in the Spanish sample). Larger correlational differences were found between form–meaning and derivatives (*r* =.89 for the Chinese and .94 for the Spanish group), form–meaning and collocations (*r* =.85 and .94 respectively), derivatives and collocations (*r* =. 74 and .89 respectively), and multiple meanings and collocations (*r* =. 80 and .93 respectively). Despite this slight differences in the strength of correlations, the two language groups show a similar pattern of very high interrelationships between the latent variables. This finding reveals the potential lack of independence between the word knowledge components, and confirms that this situation applies to both language groups.

8.5.3. Hypothesised model

The same hypothesised model was tested across both language populations. This model represented vocabulary as comprised of independent, but related, sub-constructs defined by their recall and recognition masteries (see Figure 27). The model fit proved to be good for both language samples, with a slightly better fit for the Chinese group (i.e., χ^2 *p* value, RMSEA, CFI) (see Table 84). The rest of evaluation criteria (AVE and CR in the respective chapters, Tables 50 and 77) performed similarly across both language groups, and the overall model reliability proved to be equally good for both samples (α =.90 for the Spanish and α =.88 for the Chinese group).

Table 84 Summary of goodness-of-fit statistics for hypothesised model in both groups.

Model	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA, 90% CI	SRMR
			Spanish group				
Hypothesised multidimensional model	26.426	16	.05	1.53	.99	.06 (.010-.111) *p* = .24	.013
			Chinese group				
Hypothesised multidimensional model	16.117	16	.49	1.01	1.00	.00 (.000-.068) *p* = .83	.019

This hierarchical model was designed to test the multidimensionality of vocabulary knowledge. Inspection of the internal structure of this model in both language populations (Figure 17 in Chapter 6 and Figure 29 in the current chapter) shows that the so-called dimensions of vocabulary cannot be considered independent constructs for either language group. The relationship between the latent vocabulary components and general vocabulary knowledge ranged between .94-.98 in the Spanish group. This lack of discriminant validity between the constructs was even more obvious in the Chinese group, where the regression coefficient between vocabulary knowledge and the multiple-meanings component was of 1.00 (range of all components between .82 and 1.00). These extremely high regression coefficients confirm that, statistically, the vocabulary knowledge components can only be considered one construct for L2 learners of English. Initial indication of this lack of multidimensionality was already observed in Study 3 in Chapter 6, and the present study seems to verify these results.

This absence of a multidimensional behaviour of vocabulary knowledge found for both L2 learner groups means that the alternative predictive models of vocabulary knowledge suggested in Studies 3 and 5 cannot be interpreted. If the vocabulary components do not behave as independent dimensions in the mental lexicon, any predictive relationships specified between them would be pointless. Thus, it can be concluded that no word knowledge component has the capacity of being a predictor of another word knowledge component, because they are parts of the same dimension.

To sum up, the analysis of the hierarchical multidimensional model provides the same results for the Spanish and Chinese populations: the vocabulary knowledge components are so strongly related that they do not behave as distinct dimensions in the mental lexicon of L2 learners. For this reason, the respecification of the model was needed in both cases to try and obtain a general conceptualisation of vocabulary knowledge that was supported by the data.

8.5.4. Revised model

The CFA model in which vocabulary knowledge was comprised of the eight word knowledge components operated well for both language groups. The model fit for this revised model proved to be very good for both populations, although with better RMSEA and CFI results in the Chinese group (see Table 85). This demonstrates the validity of the unidimensional vocabulary knowledge construct. Not only this model met construct validity for both groups, but also convergent and discriminant validity, as represented by the factor loading coefficients and significance levels of the paths (see Table 54 in Chapter 6 and Table 81 in the present chapter). Therefore, this revised, unidimensional model was cross-validated for both the Chinese and the Spanish learner groups, and proved to be a better empirical representation of vocabulary knowledge than the multidimensional model in these populations.

Table 85 Summary of goodness-of-fit statistics for revised model in both groups.

Model	χ^2	*df*	*p* value	χ^2/df	CFI	RMSEA, 90% CI	SRMR
			Spanish group				
Revised unidimensional model	22.078	16	.14	1.16	.99	.05 (.000-.099) *p* = .44	.013
			Chinese group				
Revised unidimensional model	16.117	16	.49	1.01	1.00	.00 (.000-.068) *p* = .84	.019

A visual inspection of the revised model for both language populations reveals a similar pattern in the behaviour of vocabulary knowledge. All the word knowledge aspects contribute very highly and almost at the same level to knowledge of vocabulary in the two groups, although the factor loadings for every word knowledge aspect are slightly lower in the Chinese group. The range of factor loadings for this group is from .67 to .88, whereas for the Spanish group it is from .81 to .93. Thus, the variance in vocabulary knowledge overall explained by these word knowledge components ranges between 44% and 77% in the Chinese group, and from 66% to 87% in the Spanish sample. Despite this somewhat smaller strength of relationship between vocabulary components and the vocabulary knowledge construct in the Chinese group, the individual components seem to follow the same trend regarding which one represents vocabulary more strongly. That is, for both the Chinese and the Spanish groups, the lowest factor loading is found in the Collocation Recognition component (R = .67 and .81 respectively), followed by the Form–meaning Recall (R = .68 and .85 respectively). Conversely, the Derivative Recall and Multiple-Meanings Recall aspects proved to yield one of the highest factor loadings for both groups. For the Chinese group, Multiple-Meanings Recall was the strongest contributor to vocabulary knowledge (R = .88), followed by Derivatives Recall (R = .84). In the Spanish group, Derivatives Recall experienced the strongest factor loading (R = .93), followed by Derivatives Recognition (R = .92) and Multiple-Meanings Recall (R = .90). This indicates that knowledge of multiple meanings and derivatives, especially in the recall mastery, explain the most variance on overall vocabulary knowledge for L2 learners. In other words, these types of knowledge seem to be the best representatives of L2 overall vocabulary knowledge.

In addition, the recall aspects seem to contribute a little more strongly to vocabulary knowledge, than the recognition aspects (except for form–meaning recall). This suggests that their interrelations with each other are more powerful, thus supporting previous claims that most complex aspects are more related to each other than easier aspects (Greidanus, Bogaards, van der Linden, Nienhuis, & de Wolf, 2004; Laufer & Goldstein, 2004; Schmitt, 2014).

The two language groups also coincide in the error correlations between the recall and recognition pairs of aspects. Only the correlation between the derivatives aspects and the collocation aspects showed to be significant for both groups, indicating that the recall and recognition masteries of those aspects assessed some common

knowledge not fully represented in the model. In the case of derivatives, this correlation explains about 8% of the variance in recall and recognition derivatives knowledge for the Chinese group, and 34% for the Spanish group. That is, there is a significant amount of general derivatives mastery that is explained by primarily non-lexical knowledge in both samples, with a stronger influence for the Spanish group. This non-lexical knowledge may refer to morphological and grammatical awareness, that is, the metalinguistic knowledge of what word classes are and how to apply morphological and syntactic rules to form new word classes in context. This morpho-syntactic awareness has been previously found to be related to vocabulary (Zhang, 2012, 2016), and is needed in order to recognise and produce derivatives. Moreover, having this metalinguistic awareness in the L1 can facilitate its acquisition process in the L2 (Nagy et al., 2006; Ramírez et al., 2013; Scott et al., 2012). Since the morphological and syntactic rules between Spanish and English are more similar than between Chinese and English, the variance explained by this knowledge in the Spanish group is much higher than that for the Chinese group.

As for the collocational knowledge, the significant error correlation seems to indicate that mastery of collocations in recall and recognition requires knowledge of the contextual situation in which they can be used, as well as developing native-like intuitions about their use (Siyanova-Chanturia & Spina, 2015). This type of knowledge influences the Chinese learners' collocational knowledge more than the Spanish learners', explaining common variance of 18% and 10% respectively.

Overall, the above analysis confirms that this unidimensional model is a better representation of vocabulary knowledge than the multidimensional interpretation for both language groups. This revised unidimensional model indicates that vocabulary knowledge is empirically one unique construct comprised of types of knowledge that have different difficulty levels. Nevertheless, the above comparison of the similarities and differences in the vocabulary knowledge conceptualisation between the Spanish and Chinese groups is based only on a visual inspection. While this can give us a general idea regarding whether vocabulary knowledge seems to behave similarly across them, a statistical, simultaneous comparison between the best-fitting model in both groups can provide a more comprehensive understanding on how the different word knowledge components and the theoretical vocabulary knowledge construct function in L2 learners. SEM software can perform multi-group analyses that statistically compare the

model results for multiple language populations. The next section reports on this analysis.

8.5.5. Multigroup CFA of the revised model

The SEM results for the Spanish group in Study 3 (Chapter 6) and for the Chinese group in the present study yielded an identical unidimensional model as the best representation of vocabulary knowledge for each group. The analyses above also suggest that there seems to be a similar trend of how the word knowledge components perform in both groups. However, the previous analyses cannot guarantee that the word knowledge measurements and the underlying theoretical vocabulary structure is equivalent across both groups. In other words, it cannot inform of whether the unidimensional vocabulary knowledge construct behaves equivalently across both language populations (Byrne, 2016). Therefore, it needs to be tested whether the vocabulary model structure differs across groups. Unlike the previous analyses which were based on single groups (i.e., Spanish and Chinese independently), to answer this question a multigroup analysis is required. In particular, this study would benefit from a multigroup confirmatory factor analysis.

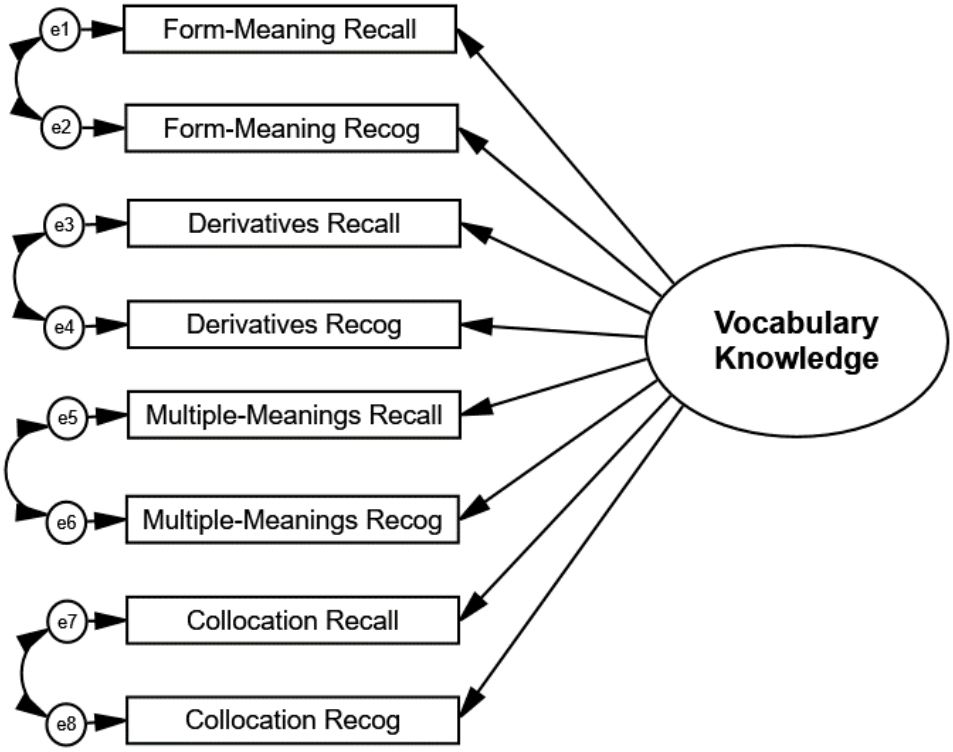
Multigroup CFA is an extension of the basic single-group CFA reported in the revised model, and is the most widely used method to test for model invariance. It does so by testing each time more restrictive cross-group models and comparing them with less restricted ones (Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). The main advantage over visual inspection and comparison of the single-sample models is that it can be used to statistically determine if and how a model differs for different language learner groups. Moreover, it analyses these relationships and potential differences simultaneously for multiple groups, providing more reliable results. For these reasons, I employed Multigroup CFA to empirically compare the results for the Chinese and Spanish samples, and evaluate whether the contribution of the different word knowledge components to the vocabulary construct differs as a function of the language background of participants. This multigroup CFA was conducted using the software AMOS 24 (Arbuckle, 2016) in SPSS 24 (IBM Corp., 2016) with the ML estimator.

Multigroup analyses can inform of model invariance at different levels. In this study I am interested in checking two main issues: 1) whether the word knowledge components that comprise vocabulary operate equivalently across the Spanish and Chinese learners, and 2) whether the theoretical construct of vocabulary knowledge is equivalent across both groups, and thus, the validity of the vocabulary construct conceptualisation confirmed. These issues require testing for group-invariance at the measurement-model stage and the latent-mean stage. In order to demonstrate measurement-model invariance, equivalence or partial equivalence is needed at various hierarchical levels: individual factor loadings (a.k.a. *weak invariance*), intercepts (*strong invariance*) and error correlations (*strict invariance*) (Kline, 2016). In order to claim latent-mean invariance, at least partial strict measurement invariance is required. To test for strict invariance, at least partial intercept invariance is needed, and, in turn, factor loading invariance is compulsory in order to test for equality of intercepts (Steinmetz et al., 2009). Therefore, the application of multigroup CFA involves the examination of each time more restrictive hierarchical models starting by factor loading invariance which, if confirmed, will be followed by intercept invariance, which, in turn, will continue with error correlation invariance once satisfied. If error correlation invariance is fulfilled, then latent mean invariance will be tested. Only if this invariance between groups is demonstrated I can claim that the construct being measured is the same in the different language groups.

Since the single-group analyses performed in Study 3 (Chapter 6) and above suggest that the unidimensional vocabulary model is the one that best fits the data for the two groups, this model was the one tested for multi-group invariance (presented schematically in Figure 33). The procedure starts with the *configural model*, which requires the same number of factors and the same factor loading patterns in both groups, although all parameters are allowed to differ across groups. The model fit for the configural model serves as reference to compare whether the potential group differences in the subsequent models are statistically significant or not (Byrne, 2016).[22]

[22] In this study I applied the most conservative practice of comparing all the models to the configural model. A more lenient approach is to compare each measurement model to the previous model instead to the original. When applying the second approach, there were not differences between the two groups at any of the measurement levels, suggesting the word knowledge components and the concept of vocabulary was completely equivalent between groups. Although the general conclusion is the same with either approach, I decided to report the most traditional procedure in order to better represent some slight differences across populations.

Figure 33 Baseline model for multigroup analysis.



After setting the configural model, I started by examining measurement invariance. This will allow me to demonstrate whether the word knowledge components in the CFA model are equivalent across the Spanish and Chinese populations. This involves a series of hierarchical steps.

First, a baseline model needs to be specified. In this baseline model, all factor loadings were fixed to be equivalent across the groups, while the latent means, indicator intercepts and error correlations were allowed to differ by group. To investigate whether the factor loadings of word knowledge components are equivalent by language group, I compared the goodness-of-fit of the baseline model with its analogous configural model, in which individual factor loadings were allowed to differ among the groups. To decide on the invariance of factor loadings between the two groups, I applied the CFI difference test (ΔCFI)[23], following the advice from multiple researchers (Byrne, 2016; Cheung & Rensvold, 2002; Kenny, 2016; Meade, Johnson, & Braddy, 2008). This is a

[23] Traditionally, the way of testing for measurement invariance has been employing the likelihood ratio test, which evaluates whether there is a significant difference in the χ^2 statistics and their degrees of freedom when applying equality constraints. However, since this statistic is very sensitive to sample size and very stringent, it has been suggested that other goodness-of-fit indexes are employed to decide for invariance of measurement tests. In particular, the ΔCFI has been claimed to be a more appropriate and practical value, since it is independent from sample size and model parameters.

value derived from the difference between the CFI of the specific constrained models and the configural model. If the ΔCFI value is smaller than or equal to (≤) -.01, this is evidence of invariance (equality) between the two groups (Cheung & Rensvold, 2002). Conversely, when ΔCFI is larger than -.01 it indicates the constrained models have a significantly worse fit than a less constrained model.

Following this, if the ΔCFI showed equivalence in the factor loadings across both groups, then the next level of invariance (i.e., intercepts) is tested. Conversely, if the ΔCFI shows a statistically significant difference between the two groups regarding the factor loadings, we need to fit subsequent and more restrictive models constraining equal one by one each factor loadings in order to identify which ones are non-equivalent. The factor loading found to differ between groups are then freely estimated, that is, allowed to differ between groups to control for these differences in the subsequent tests. Otherwise, it would be constrained equal for the subsequent models. In order to test for intercept and error correlation invariance, the same process described for factor loadings was followed.

By conducting these series of tests I determined whether the relationships between the word knowledge components and the vocabulary knowledge latent construct were the same for the two language learner populations, or whether allowing some of the relationships to differ between the two populations would significantly improve the fit of the model. The sequence of increasingly more constrained models fitted to test measurement invariance and their model fit are summarised in Table 86. Following Boomsma (2000), multiple fit indices are reported to decide on the best-fitting model.

Table 86 Summary of goodness-of-fit statistics for test of multigroup measurement invariance.

Model description	Comparable model	χ^2(df)	χ^2/df	RMSEA	SRMR	CFI	ΔCFI
1.Configural model: No equality constraints imposed	_____	43.425 (32)	1.36	.03 (.000-.057) p = .86	.014	.995	_____

Model description	Comparable model	χ^2(df)	χ^2/df	RMSEA	SRMR	CFI	ΔCFI
2. Factor Loadings *Model A*: All factor loadings constrained equal	A vs. 1	68.227 (39)	1.75	.05 (.029-.068) p = .86	.017	.987	-.008
3. Intercepts *Model B*: factor loadings and intercepts constrained equal	B vs. 1	172.547 (46)	3.75	.09 (.079-.109) p = .00	.027	.942	**-.053**
Model C: Model A with 1st intercept constrained equal	C vs. 1	68.227 (39)	1.75	.05 (.029-.068) p = .86	.017	.987	-.008
Model D: Model C with 2nd intercept constrained equal	D vs. 1	75.995 (40)	1.90	.05 (.035-.072) p = .35	.018	.984	**-.011**
Model E: Model C with 3rd intercept constrained equal	E vs. 1	68.326 (40)	1.71	.05 (.027-.067) p = .56	.017	.987	-.008
Model F: Model E with 4th intercept constrained equal	F vs. 1	104.035 (41)	2.54	.07 (.054-.087) p = .02	.018	.971	**-.024**
Model G: Model E with 5th intercept constrained equal	G vs. 1	69.073 (41)	1.69	.05 (.027-.066) p = .58	.017	.987	-.008
Model H: Model G with 6th intercept constrained equal	H vs. 1	113.214 (42)	2.70	.07 (.058-.090) p = .01	.018	.968	**-.027**
Model I: Model G with 7th intercept constrained equal	I vs. 1	77.162 (42)	1.84	.05 (.033-.070) p = .41	.017	.984	**-.011**

Model description	Comparable model	χ^2(df)	χ^2/df	RMSEA	SRMR	CFI	ΔCFI
Model J: Model G with 8th intercept constrained equal	J vs. 1	73.488 (42)	1.75	.05 (.030-.067) *p* = .51	.017	.986	-.009
4. Error correlation *Model K*: Model J with all error correlations constrained equal	K vs. 1	89.281 (46)	1.94	.06 (.038-.072) *p* = .30	.016	.980	**-.015**
Model L: Model J with 1st error correlation constrained equal	K vs. 1	73.448 (42)	1.71	.05 (.028-.066) *p* = .55	.017	.986	-.009
Model M: Model L with 2nd error correlation constrained equal	L vs. 1	87.831 (44)	2.00	.06 (.039-.074) *p* = .25	.016	.980	**-.015**
Model N: Model L with 3rd error correlation constrained equal	N vs. 1	73.693 (44)	1.68	.05 (.027-.065) *p* = .60	.017	.987	-.008
Model O: Model N with 4th error correlation constrained equal	O vs. 1	75.598 (45)	1.68	.05 (.027-.065) *p* = .60	.018	.986	-.009

Note: ΔCFI = difference in CFI values between models. Chinese group *n* = 170, Spanish group *n* = 144. Following Awang (2012) the χ^2 *p* value is not reported, as it is substantially affected by sample size. **Bold numbers** = ΔCFI higher than -.01, and thus parameters allowed to differ across both groups.

Table 86 shows that, consistent with the results of the single-group analysis, the configural model is well fitting across both the Chinese and Spanish population. This indicates that the number of factors and pattern of the factor structure is similar across both groups. However, despite this, it is possible that the factor loading behaviour is not

identical across both groups. To examine this, Model A was tested constraining all the factor loadings equal across groups. The goodness-of-fit of this model was good (see Table 86). Comparison of Model A with the configural model to determine factor loading invariance yields a ΔCFI value that suggests that the factor loadings behaved equivalently across both groups (ΔCFI = -.008). That is, factor loading invariance was fully satisfied. This indicates that the word knowledge components are manifested in the same way in each group, that is, they define and measure the same construct in each group. This does not mean, however, that the mean score of factor loading parameters needs to be exactly the same across both groups, but that together in relation to the vocabulary factor the components perform similarly in both groups. This notion was anticipated in the visual analysis, where the best and worst component contributors to vocabulary were the same in both groups.

Since factor loading measurement invariance was confirmed, then intercept invariance was tested. Intercepts are estimations of the score of an indicator, and could be interpreted as the observed mean of an indicator. Model B, in which all factor loadings and intercepts were constrained equal, was compared to the configural model. This model fit showed to be significantly worse (ΔCFI = -.053), suggesting that some intercepts behaved differently across groups. Therefore, follow-up analysis testing the invariance of all intercepts separately were performed in order to determine which one(s) performed differently (see Models C-J in Table 86). By fitting these models, I found out that the intercepts for the Form–meaning Recognition, Derivatives Recognition, Multiple-Meanings Recognition and Collocation Recall tests differed between both groups (Models D, F, H and I, respectively). This indicates that the response pattern of these word knowledge aspects is somewhat different across groups, and thus are the aspects most likely to significantly vary in their mean scores in different language populations (Kline, 2016). Conversely, the intercepts of Form–meaning Recall, Derivatives Recall, Multiple-Meanings Recall and Collocation Recognition operated the same way in the two groups (Models C, E, G and J). That is, similar-proficiency learners from the two groups would be expected to respond in the same way to those components, and thus have a similar mean score. These findings confirm the MANOVA results in Study 4 (Chapter 7), which found that the mean scores of Form–meaning Recognition, Derivatives Recognition and Multiple-Meanings Recognition

differed significantly (although not by much) across groups, with the Collocation Recall aspect approaching significance (see Chapter 7, Section 7.5.4.2).

Overall, the results of this process suggest that the unidimensional model of vocabulary knowledge was partially invariant in intercepts across the two language groups. Since general measurement invariance can exist even with partial intercept invariance (Steinmetz et al., 2009), this allows us to test the final and highest level of measurement invariance: error correlation equality.

To check whether the error correlations between word knowledge components behaved the same across groups, I compared Model K with the configural model. Model K constrained as equal all factor loadings, the 1st, 3rd, 5th and 8th intercept (thus, freeing the 2nd, 4th, 6th and 7th) and all the error correlations. The fit of Model K indicates that some error correlations seemed to behave differently across groups (ΔCFI = -.015), so I constrained them equal one by one to identify which ones. The results showed that only the error correlation between the recall and recognition derivatives seemed to perform differently in both tests (Model M ΔCFI = -.015). The final Model O with all factor loadings constrained as equal, plus the four intercepts and all error correlations except the 2nd, yielded a ΔCFI of -.009. This indicates that, despite some partial intercept and error invariance, strict measurement invariance of the vocabulary knowledge factor and its components is confirmed for both the Chinese and the Spanish populations. This strict measurement invariance implies that the word knowledge components seem to measure the vocabulary knowledge dimension similarly and with an identical degree of precision in each language group (Deshon, 2004; Kline, 2016).

The above tests have demonstrated full invariance at the factor loadings level, and partial invariance at the intercept and error covariance level. That is, partial strict measurement invariance has been satisfied, implying that the vocabulary measurement instruments employed in this thesis behave generally the same across populations. This is a requirement for testing latent-mean equivalence, which enables to demonstrated the group-invariant validity for the theoretical structure of the vocabulary knowledge construct (Byrne, 2016). Thus, as a final step, I estimated the latent-mean difference between the Spanish and Chinese groups on the vocabulary knowledge dimension. A non-significant mean estimate indicates that there is no significant difference in the latent mean of vocabulary knowledge between both groups, that is, there is latent-mean invariance. Table 87 shows the results for this test employing Model O, in which not all

measurement parameters are constrained equal across groups. Since the difference in the vocabulary knowledge latent mean across both groups was found to be non-significant (p = .89), this suggests that there seems to be little difference between the Spanish and Chinese participants in their overall vocabulary knowledge mean structure. Thus, we must conclude that there is evidence for the word knowledge component instruments and the unidimensional concept of vocabulary knowledge to be equivalent in both language populations. That is, the theoretical structure of vocabulary knowledge is the same for both groups, so that they behave similarly with respect to the vocabulary construct and its different aspects. This provides further evidence for the construct validity and representation of vocabulary as a unidimensional model.

Table 87 Latent mean difference between Spanish and Chinese learners.

Latent variable	Estimate	SE	CR	*p* value
Vocabulary Knowledge	.04	.31	.14	.89

As a summary of this long process, Table 86 shows that the multigroup configural model fit is acceptable, confirming that the vocabulary construct has the same number of factors and patterns across both groups. The second model, measuring factor invariance, also yielded an acceptable fit, suggesting that the construct has the same meaning in the two groups. The model for full intercept invariance (Model B), however, showed to be significantly worse than the configural model, and thus four intercepts had to be allowed to differ in order to obtain partial intercept invariance. The same was the case for the error correlation invariance (Model K), with one error correlation being freed in the model. I finally measured full latent mean invariance, and the results showed that both L2 learner groups seemed to have a similar overall vocabulary latent mean. Therefore, this multigroup CFA comparison demonstrates that the word knowledge components and the unidimensional vocabulary knowledge construct is equivalent across both language populations.

8.6. Discussion

The analysis of the nature of vocabulary knowledge in Study 3 (Chapter 6) and the present study has involved: 1) testing the validity of the hypothesised multidimensional model of vocabulary knowledge for the Spanish and Chinese groups separately; 2) giving findings on the inadequate fit of that model, and the proposal and testing of an alternative unidimensional model; 3) cross-validating this model for each language group; and 4) assessing the equivalence of this vocabulary knowledge model across both language populations. In this section, I will summarise the general findings for the Chinese and Spanish L2 learners regarding the relationships between word knowledge components, and between the components and the vocabulary construct[24].

8.6.1. Behaviour of word knowledge aspects and the vocabulary knowledge construct in L2s

As discussed in Chapter 2, different categorisations of aspects involved in knowing vocabulary present them as independent concepts. Although the authors of these classifications do not intend to convey the message that the aspects are not related, research has generally treated them separately (e.g., Henriksen, 1999), which has limited our understanding of how they relate to each other. Yet, there have been claims that the components are interrelated in some way (e.g., Qian, 2002; Vermeer, 2001; Webb, 2008), and many researchers and practitioners would agree about their interconnections. The present study confirms that the word knowledge aspects are highly interrelated with one another for L2 English learners of different L1s. These findings indicate that word knowledge aspects cannot be interpreted as anything other than correlated components in the mental lexicon of L2 learners, regardless of how different the L1 is from the L2. Still, the intercorrelations (as demonstrated by Spearman's rho and factor loadings) among vocabulary components for the Spanish population are higher than those for the Chinese sample. In a study employing SEM, Kieffer and Lesaux (2012c) found that two lexical aspects (i.e., morphological awareness and reading vocabulary) were slightly more strongly interrelated for a group of Spanish learners of English than for a group of Vietnamese learners of English.

[24] The discussion section in this chapter complements that of Chapter 6. A more detailed interpretation of specific results may be found in that chapter.

Although the differences were not very large (just as in my studies), these findings might indicate that greater similarities between the L1 and L2 may be reflected in higher interrelationships among vocabulary aspects.

Some differences between the two L2s were also found regarding their performance in the tests, with the Chinese participants overall showing slightly less consistency in their responses. Again, this might be a reflection of the L1-L2 similarities. The language commonalities between Spanish and English mean that, in answering the vocabulary knowledge tests, these learners can make more connections between their L1 and their L2, facilitating more consistency in their answers. Conversely, the Chinese language shares less features with English, and thus the learners cannot rely as much on their L1 knowledge or language similarities to complete these tests. Hence, their L2 mastery would mainly depend on their different experiences with the language, which would reflect more individual differences and, as a result, more variability in the answers.

Nevertheless, this study displays more similarities than differences in the overall behaviour of vocabulary knowledge in L2s. As anticipated in Chapter 6, the present study corroborates the finding that vocabulary knowledge does not behave as a multidimensional construct in the L2 learners' minds. The word knowledge components specified as different dimensions did not achieve discriminant validity for either the Spanish or the Chinese learner groups, providing further evidence that they cannot be interpreted as independent dimensions. The Spanish and Chinese languages are not only very different from each other, but also in their connection with the English language. This lack of support for the multidimensionality of vocabulary knowledge in L2 learners of two very diverse linguistic backgrounds provides robust evidence for the strength of this result. As a consequence, it seems reasonable to conclude that vocabulary knowledge is not multidimensional in second language acquisition.

What is the best way of conceptualising vocabulary knowledge, then? The results from the present study reinforce the findings from Study 3 (Chapter 6) that vocabulary knowledge should be best considered a unidimensional construct. This construct is, in turn, represented by various difficulty levels of knowledge, which are the traditional recall and recognition word knowledge aspects. Support for a unidimensional understanding of vocabulary knowledge being evidenced by two very different language populations suggests the generalisability of the results. From these

findings it follows that the behaviour of overall vocabulary knowledge in the L2 learners' mental lexicon is the same regardless of the learners' L1. Previous studies have found that overall vocabulary knowledge seems to function similarly in different groups of L2 English learners. For example, Kieffer and Lesaux (2012a) discovered that multiple groups of L2 English learners with L1s as different as Spanish and Vietnamese exhibited similar evidence for the contribution of vocabulary components to the general construct, as well as in the theoretical understanding of the concept. My findings support the conclusion that the theoretical conceptualisation of vocabulary knowledge seems to be the same for different L2 learners, and shows that this conceptualisation behaves in a unidimensional manner.

This general unidimensional conceptualisation of vocabulary knowledge should not be seen, however, as completely breaking with previous research on vocabulary. Rather, it should be viewed as an empirical framework that explains earlier findings. One of these is the relationship between depth and breadth of vocabulary. Although generally treated as independent constructs, much earlier research has shown high intercorrelations between them (e.g., r = .88, Tannenbaum et al., 2006), which have led some researchers to recently claim they are not separate constructs (Gyllstad, 2013; Kieffer & Lesaux, 2012a; Schmitt, 2014). My study provides empirical support that size and depth, although a useful distinction for vocabulary teaching, cannot be considered as independent and completely distinct from each other. It does so by demonstrating not only high interconnections between each individual word knowledge aspect, but also a lack of practical discrimination between them. This suggests that multiple, varied and rich experiences with diverse words not only lead to basic knowledge of more words (i.e., size), but also to deeper knowledge of each word in different contexts and meanings (Tannenbaum et al., 2006). Overall, the findings suggest that the development of word knowledge components occurs concurrently, which supports earlier claims that improvements in one aspect lead to improvements in others in a somewhat parallel trend (Schmitt, 1998; Webb, 2005).

8.6.2. Equivalence of the vocabulary knowledge construct in L2 learners

According to Bachman (1990), the description of linguistic constructs involves three steps: first, describing it theoretically, then, operationally, and finally, establishing

procedures to quantify the observations from those constructs. The first stage requires the description of the particular features of the type of knowledge we are measuring, including the relationships between its sub-constructs, if applicable. In the case of the present study, this theoretical description was taken from Nations' (2013) word knowledge framework. The second stage involves developing an appropriate measurement instrument to make the construct observable. This is represented by the word knowledge components test battery developed in Chapter 3 (also see Appendix 3). The third stage requires quantifying the tests into a scale, which was done as part of the implicational scaling analysis and in the evaluation of the hypothesised multidimensional vocabulary model in Studies 1, 3, 4 and 5.

Yet, in order to prove the generalisability of a construct, an additional stage is required: the cross-validation and invariance of the construct in different language populations. The findings of the multigroup CFA analyses revealed that, indeed, the unidimensional vocabulary construct is empirically valid and operates equivalently across at least two different language populations. The independent word knowledge components were shown to measure the same concepts and perform similarly across both groups. This implies that when learners of different L1s encounter these components, their performance in each of them would follow the same pattern relative to each component. This is true despite the cognate effect on most components found for the Spanish participants in Study 2 (Chapter 5). These findings seem to indicate that when learners have a similar vocabulary mastery, the possible cognate status of their L1 does not significantly modify their overall behaviour in the different word knowledge aspects when compared to learners from non-cognate backgrounds.

While the role of the word knowledge components in relation to each other has been demonstrated to be generalisable across both language groups, some aspects were found to reflect certain group-related influences. These were the recognition of form–meaning, derivatives and multiple meanings, and the recall of collocations. This finding suggests that these aspects are the most likely to vary in their functioning in different language groups. Since the components most influenced by these group-specific features were generally recognition-based aspects, they might be revealing the effect of alphabetic differences between the two language groups when encountering a word in reading. That is, learners whose L1 and L2 alphabetic systems share common features would have a general advantage in the recognition of knowledge, as suggested by Webb

and Nation (2017). In the case of collocation recall, the contextual differences between Chinese and Spanish with English might have influenced the pattern of responses to this context-based aspect. Due to these differences, Chinese learners might be expected to share less contextual information with the English language than the Spanish language. As a case in point, the Chinese learners as a group had trouble answering the collocation *grated cheese*, regardless of their proficiency level, because 'cheese' is not common in the Chinese diet. Conversely, this concept is more usual in Spanish gastronomy, and therefore the correct response to this collocation depended on each Spanish learner's vocabulary proficiency rather than on cultural/contextual differences. Another feature that varied slightly between the Spanish and Chinese populations was the error correlation[25] between the two derivative tests. This variation indicates different effects that knowledge of morphological and syntactic rules has on each language group. While in the Spanish group this influence seemed to explain about 34% of the common unexplained variance in derivatives knowledge, it only explained 8% in the Chinese group. Thus, the Spanish language's morphological and grammatical similarities with English seem to lead to a higher morpho-syntactic awareness, and thus, a greater influence in derivatives knowledge.

Nevertheless, this group-specific influence is not strong enough to have affected the behaviour of the word knowledge in general, and the similarities between the two groups' vocabulary performance are more apparent. Hence, overall the multigroup analysis revealed that general measurement invariance was found between the Spanish and the Chinese samples, indicating that the word knowledge components, in relation to the vocabulary construct, operate in a similar trend across both groups. That is, the functioning of the general vocabulary construct was found to be invariant across both learner populations.

To sum up, my study shows that the vocabulary knowledge structure is equivalent in the two language groups, indicating that it represents the same notion in the mental lexicon of L2 English learners. This suggests that the unidimensional conceptualisation of vocabulary knowledge represented by the eight word knowledge components could be a universal theoretical description of the construct.

[25] As explained in Chapter 6, error correlations suggest common variance between two aspects not fully explained in the model.

8.6.3. Implications for understanding and measuring vocabulary knowledge

The great complexity of vocabulary knowledge makes this construct fundamentally difficult to conceptualise and measure. A complete analysis of this construct would concern various and varied issues, such as the number of words known (including sampling which words to test and how many), the depth in which those words are known (in terms of different types of knowledge), the level to which they are known (recall vs. recognition, receptive vs. productive), and the mode in which they can be used (spoken vs. written). This intricacy results in an impossibility of acknowledging all (or many) of these matters in one study, limiting our knowledge of the construct and leading to the lack of a general, empirically-sustained understanding of overall vocabulary knowledge.

The present study, although unable to address all of the above issues, provides the most comprehensive analysis of general vocabulary knowledge to date. Thus, the results have implications for the interpretation of vocabulary knowledge and its measurement. My findings suggest that vocabulary knowledge is not multidimensional in practice, and thus considering it so may lead to inappropriate conclusions of how the vocabulary knowledge of learners behaves in the lexicon. I found robust evidence that the vocabulary tests, although designed to theoretically reflect different dimensions of vocabulary, in fact tapped into the same and unique dimension in the mental lexicon of learners. This indicates that the word knowledge components, although hypothetically different, do not seem to represent completely separate concepts in practice. As a result, if researchers base their interpretation of these components only on theoretical grounds, they may make the mistake of treating the same underlying ability represented in various levels of difficulty as different concepts. Therefore, when measuring vocabulary knowledge in its multiple aspects, researchers and practitioners must be aware that they are tapping into the same practical dimension, but different difficulty levels of that dimension.

These findings do not mean, however, that the distinction of word knowledge components is not acceptable as a descriptive framework. In fact, this is still useful in vocabulary teaching. The various word knowledge components represent diverse levels of vocabulary mastery, and thus they complement each other. Hence, due to being part of the same underlying dimension, the results in one aspect may give some indication of knowledge of another aspect, although they cannot replace each other. My research

does not imply, therefore, that assessing only one type of word knowledge is enough to get an estimation of the learners' vocabulary mastery. In fact, it is quite the opposite, since my findings show that the different aspects of word knowledge are important for the understanding of vocabulary knowledge as a whole. If we measure only one aspect, the results would be influenced not only by the overall vocabulary mastery of a learner, but also by the specific type of knowledge and its difficulty. Thus, my results emphasise the need to use comprehensive assessment of vocabulary knowledge. Practitioners and researchers seeking to obtain a more robust and complete understanding of the L2 learners' vocabulary knowledge construct must measure multiple levels of this dimension, since each aspect can provide a different insight into this knowledge.

My study also points to the need for the consideration of some terminology used in previous vocabulary studies to refer to the construct, which might have misled our interpretation of vocabulary knowledge in the past. The main issue is related to the nomenclature of the dimensions approach, which implies, by the use of the term *dimensions*, that the various word knowledge aspects are fundamentally distinct constructs in vocabulary knowledge (e.g., Read, 2000). Although many researchers indeed argued for the interrelationships between these aspects, they have been typically interpreted and treated as separate dimensions, spreading this multidimensional understanding in the field. My findings show that they cannot be considered essentially independent dimensions (in the strict sense of the term, see Chapter 6, Section 6.6.5), but aspects of the same unique construct. Thus, the use of the above terminology can lead to misconceptions regarding our conceptualisation of the relationships between word knowledge components in the mind.

Finally, the findings from the current study suggest that L2 learners from different linguistic backgrounds demonstrate the same overall behaviour of the word knowledge aspects and the concept of vocabulary in their minds. Nevertheless, some specific aspects were shown to behave less similarly across both groups than others, suggesting some group-specific influence. Researchers and practitioners should take this into account when assessing vocabulary knowledge in different learner groups, as a detailed interpretation of this knowledge should not be based on only one test which may be reflecting L1-based factors. In order to enable a robust comparison of vocabulary knowledge between samples, assessing knowledge of multiple aspects

would provide a better representation of the general vocabulary mastery of different groups or learners.

8.6.4. Pedagogical implications

Certain findings of the present study can inform the vocabulary pedagogical practice. The first one refers to the demonstrated importance of the distinction, in terms of difficulty, between recognition and recall knowledge. When the recall and recognition masteries of one aspect were combined into one factor in the multidimensional model, the results for both language groups showed a significant discrepancy in the data. Therefore, it appears that the recall (productive) and recognition (receptive) mastery is a key distinction in vocabulary knowledge.

The recall/recognition distinction can have several implications for vocabulary teaching, depending on the purpose of instruction and the proficiency level of the students. If the teaching goal is to get our learners to only receptive mastery, where, for example, they just need to understand readings, then my findings suggest that teachers should focus on presenting receptive knowledge of multiple aspects. In this case, learners can make progress on their receptive knowledge of the different components without having to rely on any productive knowledge. Conversely, if attaining productive vocabulary knowledge is the goal, then the focus of pedagogy should be shifted towards pushing learners' knowledge from receptive to productive mastery. Language teaching materials, if they focus on vocabulary at all, generally merely introduce new words, often as glosses of reading passages or as defined words in vocabulary boxes. The lexical descriptions or tasks are usually limited to the meaning of a word, aiming to help the learner recognise and understand the word in discourse (i.e., receptive knowledge). While this is necessary to develop receptive mastery, there is typically very little recycling of vocabulary, or the kind of exercises necessary to develop productive mastery of the word (Brown, 2011). Thus, if productive mastery is the target, vocabulary needs to be taught through productive learning activities (Schmitt, 2008; Webb, 2009). My research shows that this productive mastery would require more effort, time, and background knowledge to be developed, since it is not to be expected at least until various aspects are mastered in recognition.

This takes us to considering the level of our students for the application of these findings. For beginner learners, my study shows that a teaching focus on receptive

mastery is probably the most appropriate. This will provide them with the essential knowledge to start understanding the language, along with the necessary background to make the move to productive knowledge. These learners should not be expected to produce vocabulary until they have the necessary receptive knowledge of multiple aspects to make the transition to production. For intermediate and advanced learners, knowing aspects at a productive level is needed, and teachers should concentrate on providing extra practice and attention on productive activities to help these types of students achieve this mastery.

The various word knowledge components also proved to be essential for vocabulary knowledge, and thus should be a focus in teaching. The general implication from my study is that, in order to enable students to use the words well, robust vocabulary instruction should focus on enhancing every aspect of word knowledge, not only the form–meaning link, including activities that present words in various types of knowledge and situations (Beck et al., 2013; Spencer et al., 2015). The form–meaning link recognition aspect is the most useful component to start learning/using a language, and thus should be given initial attention in the classroom. Nonetheless, given that this aspect has been found to be one of the best learnt ones in an L2 (see Studies 1 and 4), the emphasis on form–meaning recognition should be moved relatively quickly to other components that better represent overall lexical knowledge, such as collocations or derivatives.

Following the unidimensional model, the best descriptors of vocabulary knowledge were found to be knowledge of derivatives and multiple meanings, both in recall and recognition. The awareness and knowledge of these more complex components will not only improve the overall vocabulary mastery of the learner, but also help expand the more basic form–meaning mastery. Thus, quality vocabulary instruction should not overlook knowledge of derivatives and multiple meanings. Yet, these aspects proved to be very difficult, and thus they cannot be expected to be learnt until late in the learning process, after a great deal of time and plenty of exposure to the language. This is partially because they imply mastery of other aspects within them (e.g., form–meaning link). Therefore, their teaching requires serious attention. As stated in Chapter 7, derivative knowledge would benefit from a more explicit teaching focus, while greater benefits to the multiple-meanings aspect would result from extensive exposure to words in multiple and varied situations and contexts. Overall, given their

difficulty and time required to be learnt, if we want our students to get to recall knowledge of multiple meanings and derivatives, their engagement with the English language, both inside and outside the classroom, needs to be considerable.

Moreover, the unidimensional conceptualisation of vocabulary demonstrates that when we consider the different word knowledge components, we are presented with different levels of difficulty of the same construct. For practical research and pedagogical purposes, these levels of difficulty are probably better shown in isolation. Nevertheless, it is important to remember that they are very interrelated parts of a unique construct, and thus links between them need to be explicitly made during instruction.

As a way of improving the general teaching experience and ensuring good vocabulary knowledge, the teacher should first develop the learners' awareness of the different aspects involved in knowing vocabulary and their importance for communication. This is consistent with the notion of *word consciousness*, defined by Scott et al., (2012, p. 169) as a "cluster of rather diverse types of knowledge and skills that lead to an awareness of words and flexible engagement with their use". In a three-year treatment study, the authors showed that word consciousness was an effective technique, first, to build teachers' awareness of the importance of vocabulary instruction and what it should include, and second, to make learners aware of what knowing vocabulary involves, which led to better learning and accurate use of their vocabulary in various contexts.

In order to develop word consciousness that leads to using vocabulary appropriately in any situation, Scott et al. (2012) suggest vocabulary instruction based on motivational approaches. Tseng and Schmitt (2008) already found that vocabulary knowledge was the product of multiple motivational vocabulary strategies, and Dóczi and Kormos (2016) demonstrated that engaging tasks are more beneficial to vocabulary growth in ESL and EFL learners of English. With this motivational framework, the knowledge of words can be enhanced by engaging the learners in reflecting on the use of words, and developing generative activities that encourage learners create new words from words they already know. Also, in the case of Spanish learners, employing their cognate knowledge to investigate and guess the meaning of words would provide a motivational atmosphere that would enhance their vocabulary. Thus, Scott et al. (2012) suggest that a systematic and theory-based vocabulary training programme needs to encourage vocabulary awareness at a three levels: 1) *Metacognitive*, by raising

awareness of how words are acquired, used and why they are important; 2) *Metalinguistic*, by creating connections between languages with common roots, such as cognate recognition or common morpho-syntactic rules; and 3) *Affectional*, by employing games and creative activities to create motivating contexts that promote learning. The combination of these three levels of awareness in learners could lead to an augmented motivational period (what Dörnyei, Ibrahim, and Muir (2015) call *directed motivational currents*) to acquire multiple types of vocabulary knowledge which could seem unnecessarily overwhelming, and even impossible to master, without the emphasis on this vocabulary awareness.

8.6.5. Limitations and directions for future research

As a replication of Study 3 in, all the limitations discussed in Chapter 6 are still applicable to the present study[26]. Yet, there are some features specific to this paper that limit my findings and could be the foundations for future research.

First of all, the generalisability of the results are based on only speakers of two languages. Although the method of analysis employed is sophisticated and can reliably inform of this generalisable pattern, further research is needed that replicates the results with English learners of other mother tongues in order to corroborate these findings. For example, the generalisation of results to cognate languages is based on only the analysis of Spanish data. It would be interesting to test whether the findings can be reproduced for other cognate languages. Similarly, non-cognate languages include a broader variety of languages than the cognate group does. Therefore, other languages than Chinese need to be tested to determine the universality of these results. It would be expected that the results are corroborated, given the reproducibility of findings for languages as different as Spanish and Chinese. Nevertheless, this is just the initial results, and thus the universality of a unidimensional understanding of vocabulary would require further investigation.

Second, the divergence of my findings from previous studies conducted following different conceptualisations of vocabulary knowledge (Kieffer & Lesaux, 2012a; Scott et al., 2017) raises questions about how the diverse interpretations and frameworks of vocabulary knowledge affect the results on its structure. When the

[26] In order to avoid repetition, refer to Chapter 6 Section 6.6.6 for more details.

vocabulary measures are reading-based and/or designed to test knowledge of native speakers as well as non-native speakers of a language, vocabulary knowledge seems to emerge as multidimensional. Further investigation is needed to check whether different tests directed towards specific skills (e.g., writing, speaking or listening) would lead to a different conceptualisation of vocabulary knowledge. Moreover, it might be that L2 learners in the L1 environment show a different behaviour on overall vocabulary knowledge than L2 learners in their home environment (i.e., the sample in this study). Although this is admittedly only a speculation, additional examination is needed to demonstrate the effect of the learning setting.

Third, the unidimensional conceptualisation of vocabulary knowledge is based only on written-mode vocabulary measures, and therefore it prevents us from knowing whether this model generalises to the learners' oral vocabulary. It may be that the word knowledge components in this written mode were more intercorrelated and less different than they would be had I measured the word knowledge components orally. Although these findings still represent the most comprehensive view of vocabulary knowledge, future research should investigate the impact of oral-mode word knowledge tests on the dimensionality of vocabulary knowledge. In addition, my vocabulary measures assessed recall and recognition knowledge of the different components. Nevertheless, it is difficult to know whether the results can be applicable to the *use* of vocabulary in the language. Although the recognition results can inform somehow of the ability to use this components in reading, it cannot tell us much about the listening mastery; likewise, the recall findings do not inform of the ability to speak in the language, and only marginally about predictive writing skills. Therefore, future research could replicate this study with various word knowledge test formats that best represent the actual use of vocabulary.

Finally, the multigroup CFA analysis in the present study can only inform of the vocabulary knowledge of L2 English learners, and thus it would be interesting to replicate this study with native speakers of English. It is possible that a vocabulary model is found that applies to groups of L1 and L2 learners, as one previous study discovered (see Kieffer & Lesaux, 2012a). However, this study also found some models that applied only to the different groups of L2s and not native speakers. Thus, even if a unique conceptualisation of vocabulary knowledge can be found across both groups, it is likely that there are differences between L1 and L2 learners, and thus another model

is found that applies to only L1 or L2 learners. For example, Babayiğit (2015) found that L1 and L2 English learners differed in oral language processing skills, including vocabulary production, with L1 learners showing an advantage. Nevertheless, she did not find such advantage for word reading skills (i.e., recognition of words' form). Therefore, it can also be that recall components differ more between L1 and L2 learners than recognition components. Investigating the unidimensional model suggested in the current study across multiple L1 and L2 learners can provide more robust basis for the validity of a general theory of vocabulary acquisition.

8.7. Conclusion

Study 3 in Chapter 6 in this thesis investigated the dimensionality of vocabulary knowledge based on Nation's (2013) framework of what it means to know a word, and reported on innovative findings which suggest an empirical, unidimensional conceptualisation of vocabulary knowledge. Given the effect of such findings, the present study replicated the design and analysis with a different sample of L2 English learners: Chinese speakers. The findings corroborated the lack of multidimensionality of L2 vocabulary knowledge supported by theory, and verified the validity of the unidimensional interpretation of vocabulary knowledge as the best empirical representation of the construct for different L2 learners of English. Based on simultaneous statistical comparison of the Spanish and Chinese learners' results, this study also demonstrates that the individual word knowledge aspects, as well as the concept of general vocabulary knowledge, behave equivalently in the mental lexicons of diverse L2 learner populations. Overall, this replication study has helped set the basis for a general, empirically-based and comprehensive model of L2 vocabulary knowledge.

Chapter 9

General discussion and conclusion

The aim of this chapter is to contextualise the findings from the five studies reported in this thesis within the framework of previous research on depth of vocabulary knowledge. These studies combined provide a thorough examination of overall vocabulary knowledge, describing the order of acquisition of word knowledge components, the factors that influence learners' knowledge of them, and the conceptualisation of their relationships in the mental lexicon. The main findings of each study will be summarised independently first (Study 1 in Chapter 4 and Study 3 in Chapter 6 will be interpreted along with their replication studies in Chapter 7 (Study 4) and in Chapter 8 (Study 5), respectively) and then they will be discussed in relation to the others in order to provide directions for future research.

9.1. Main findings and implications

9.1.1. Order of acquisition of vocabulary components

The first study of this thesis investigated the existence of an order of acquisition of vocabulary components by L2 English learners. In doing so, an extensive test battery assessing recall and recognition knowledge of the form–meaning link, derivatives, collocations and multiple meanings was employed to obtain an estimation of the participants' degree of knowledge in each component. Results showed that Spanish EFL learners as a group knew at least half of the items in each of the recall and recognition components. Although it is likely that most of the knowledge of complex aspects such as derivatives or multiple meanings comes from the more advanced learners in the sample, this finding is encouraging insofar that it indicates that L2 learners can achieve reasonably good knowledge of even derivatives and multiple meanings with enough exposure to the language.

The findings also point out that, as much previous research has demonstrated, the participants' recognition knowledge is better than their recall knowledge (e.g., Laufer & Goldstein, 2004; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Webb, 2008). However, this study goes further and illustrates that this is not only true of every component individually (e.g., form–meaning recognition being better known that form–meaning recall), but of all vocabulary components tested. Following Read's (2000) approach that recognition tests are typically closer measures of receptive mastery and recall formats of productive mastery, this finding suggests that receptive knowledge of all aspects is greater than productive knowledge of any aspect. More importantly, Study 1 in Chapter 4 illustrates that recognition knowledge of all the individual word knowledge aspects develops before recall knowledge of any aspect in an implicational scale. This result implies that receptive knowledge needs to be mastered before productive knowledge can occur, confirming previous intuitions of the continuum relationship between receptive and productive knowledge (Melka, 1997; Nation, 2013; Read, 2000). Given these findings, it seems that the distinction between and the progression from receptive to productive knowledge is a crucial element in the development of general vocabulary acquisition.

Acquiring that productive mastery, however, showed to be problematic and to require considerable time and effort. Although it is not a new conclusion that recall knowledge is difficult (Cheng & Matthews, 2018; Pellicer-Sánchez & Schmitt, 2010; Van Zeeland & Schmitt, 2013), an original finding from this study is that the learning distance from recognition mastery to recall mastery seems to be much larger than from lack of knowledge to recognition mastery. Both the average mean scores and the number of people mastering the recognition aspects overall are much greater than the recall aspects' scores and the number of learners that mastered the recall aspects (see Section 7.6.4 in Chapter 7 for a more detailed discussion). Thus, if the aim of vocabulary instruction is for learners to use the words, pedagogical approaches should move from mainly (or only) focusing on receptive knowledge to help learners progress along the continuum by emphasising the instruction of recall knowledge and development of productive skills (Webb, 2005, 2009).

The implicational scale also provides, for the first time, an order of acquisition of individual vocabulary components that is consistent across learners of different proficiencies. The form–meaning link recognition is presented as the earliest aspect to

be mastered by learners, on which the rest of aspects develop. In their multicomponent treatment study, Chen and Truscott (2010) found recognition of the form–meaning link to appear after other aspects, suggesting that the order of individual components vary when specific treatment activities are employed for the acquisition of vocabulary. However, when all learning conditions and situations are accounted for, the form–meaning link seems to be the first type of knowledge learners rely on when acquiring vocabulary. The scale also illustrates that, although recall collocate knowledge is complex, it seems to be grasped before learners can recall derivatives and multiple meanings of words. Indeed, very few learners mastered knowledge of derivatives and multiple meanings, supporting previous claims that they are the last aspects to be acquired by learners (Jiang, 2000, 2002; Wolter, 2009). This finding suggests that derivatives and multiple meanings should be targets in vocabulary instruction for advanced learners.

Study 1 (Chapter 4) also shows that word knowledge aspects are very strongly interrelated with each other. This finding is important because it indicates that word knowledge components are not learnt in an isolated manner, but rather develop in parallel to a certain extent, providing support to previous claims (Schmitt, 2010; Webb, 2007a). If word knowledge components progress in parallel, learners would benefit from concurrent exposure to multiple aspects where knowledge gained in one component would help gaining knowledge in another aspect.

Finally, all these findings described for Study 1 were replicated for a very different population of EFL learners in Study 4 (Chapter 7). The Chinese learners, even with a general lower vocabulary knowledge, presented the same order of acquisition of word knowledge aspects, providing further evidence for what seems to be a natural order of acquisition of vocabulary components. This finding is novel in the vocabulary field, and indicates that the acquisition of lexis, like that of syntax, morphology or phonology, might be determined by certain developmental stages. Drawing on this finding, Pienemann's (1998, 2005) Processability Theory (which suggests that L2 learners develop in language following a series of hierarchical and implicationally ordered procedures) seems to also apply to the acquisition of L2 vocabulary knowledge. Furthermore, according to the Lexical Quality Hypothesis (Perfetti, 2007), during this acquisition process learners require multiple and varied experiences with the words to develop knowledge of the various vocabulary components and to progress along the

developmental sequence. This finding has practical implications by providing a model that allows us to roughly imply the vocabulary knowledge of a learner given their position on the scale. For example, demonstration of mastery of derivatives recognition can provide teachers with an indication that the learner has mastered the recognition of the form–meaning link and a frequent collocate. Yet, this additional, implied knowledge is not verified, and neither is whether the learner has more advanced knowledge than the one tested. Thus, although the scale can give a general estimation of the mastered vocabulary aspects based on the performance of learners on another aspect in the scale, measuring knowledge of each component would provide a better approximation of actual vocabulary knowledge.

9.1.2. Factors influencing vocabulary acquisition

Building on the order of acquisition presented in Study 1 (Chapter 4), the second study investigated factors that might have had an effect on the learners' knowledge of vocabulary components, and hence their acquisition order. Research has pointed towards various inter- and intralexical factors that seem to affect vocabulary learning, and Study 2 in Chapter 5 focuses on four of these features that are of particular interest for the present thesis. These are cognateness and proficiency as interlexical factors and word length and word frequency as intralexical factors. The influence of these linguistic features has generally been demonstrated primarily or uniquely for the form–meaning link aspect, in some cases with inconclusive results (e.g., word length effect). Therefore, Study 2 investigates the potential contribution of these factors to knowledge of other vocabulary components.

The lemma frequency proved to have a significant contribution on the knowledge of all word knowledge components, except collocational knowledge. This adds to the body of research demonstrating its influence on the form–meaning link and its inconclusive effect on collocational knowledge, while expanding the frequency effect to aspects such as recognition and recall of derivatives and multiple meanings. Although some research had found correlations between frequency and derivatives (Saigh, 2015) and frequency and multiple meanings (Crossley et al., 2010), its direct, main effect had not been demonstrated on these aspects, to the best of my knowledge. There seems to be two plausible explanations for the word lemma frequency affecting

knowledge of the multiple senses. On the one hand, the fact that the words' frequencies were retrieved from a corpus that is semantically untagged means that the lemma frequency count includes the frequencies of the word in its multiple meanings. As a result, this frequency would predict the knowledge of different meanings. On the other hand, it can be that words that are more frequent in their core meaning, are also more frequent in their other meanings, and thus more likely to be known in their various senses. In the derivatives case, however, only the second explanation seems reasonable, since the lemma count does not account for the frequencies of the different derivative forms. In this sense, more frequent lemmas would imply more frequent derivatives of a word. These findings regarding the frequency effect must be interpreted in light of their potential implications for teaching and material development. Although a word's various derivatives and meanings occur in language with different frequencies, my findings indicate that knowing the lemma frequency of a word can estimate the likelihood of knowing not only the form–meaning link of that word, but also its different derivatives and multiple senses. Thus, the lemma frequency can be employed as a general predictor of the probability of knowing various word knowledge components, and thus informs the selection of words to be taught in multiple meanings and derivatives as well, without necessarily having to retrieve the individual frequencies for each derivative and sense.

Study 2 (Chapter 5) also showed that proficiency was the only factor influencing knowledge of all the vocabulary components tested. Research had previously found the proficiency effect on various vocabulary components, but never in a multi-component, multi-test study. This finding is important insofar that proficiency can inform about the likelihood of knowledge of multiple vocabulary components, despite not being the strongest predictor. Since proficiency in this study was conceptualised as knowledge of the VLT, the results also seem to indicate that the vocabulary size of learners is a good indicator of general vocabulary knowledge.

Conversely, word length does not seem to be a factor that has a significant influence on vocabulary knowledge overall. The only aspect for which an effect of word length was found was, expectedly, recognition of the form–meaning link. However, surprisingly, longer words seemed to be recognised better, which contradicts most previous acquisition research on the topic (Barclay, 2017; Campoy, 2008). This result indicates that even though shorter words may be more easily learnt (as demonstrated by

Barclay, 2017; Campoy, 2008), longer words may be more easily retained, and hence once words are known, the longer ones are more salient and thus better recognised (e.g., Peters, 2016).

More importantly, the results in this second study show cross-linguistic influence, demonstrated in the generally better knowledge of cognates over non-cognates across most vocabulary components. This positive effect of cognateness status was confirmed for the form–meaning link at both sensitivity levels (i.e., recognition and recall), as much previous research has shown (e.g., Elgort, 2012; Laufer & Ravenhorst-Kalovski, 2010; Méndez Pérez et al., 2010; Rosselli et al., 2012), and also extended to general knowledge of multiple meanings and recognition of derivatives. This demonstrates that the power of a word's cognateness status extends to deeper word knowledge aspects, and thus, L2 learners can know a lot about cognate words. This cognate knowledge is largely moderated by the proficiency of learners, though, showing to be more obvious in lower-proficiency learners.

However, the cross-linguistic influence showed in the knowledge of most vocabulary components raised the question of whether the order of acquisition suggested in Study 1 (Chapter 4) represented only cognate languages. This was addressed in Study 4 (Chapter 7) by replicating the findings in Study 1 with Chinese EFL learners, and the fact that the same order of acquisition remained seems to indicate that the cognateness status of the Spanish language may help learners master the different components before learners of non-cognate languages, but does not alter the order in which they appear. Given that cognateness awareness seems to enable the transfer of knowledge from the L1 to the L2 across many aspects of vocabulary knowledge, as well as quick and efficient vocabulary growth (Dressler et al., 2011; Ramírez et al., 2013), raising awareness of the shared knowledge between cognate words in different languages should be a priority in vocabulary instruction.

Moreover, cognateness status proved to be the strongest predictor of vocabulary knowledge, over and above frequency. This is interesting to discuss in light of estimations of different populations of learners' English vocabulary knowledge. The finding seems to suggest that, in L2 English learners of cognate languages, cognateness status would be a better predictor of the likelihood of knowing a word than frequency. This has important implications for vocabulary testing. The inclusion of cognates in frequency-based tests designed to estimate the vocabulary size/level of learners could

overestimate this knowledge if the proportion in which they are included is higher than that in the two languages (Elgort, 2013). Therefore, for cognate languages, cognateness and frequency should be combined as main indicators of lexical knowledge in test development. This would provide more tailored and accurate vocabulary tests that control for the frequency of words, and also for their cognateness status in the evaluation of vocabulary sizes.

Finally, collocational knowledge was not influenced by the node's lemma frequency, word length or cognateness status. This is further evidence for the holistic nature of collocations as opposed to single words (Wray, 2002, 2008). When only the information of one of the constituents of a collocation is accounted for (either frequency, length or cognateness), no influence of such features is found, since it does not represent the characteristics of the whole collocation. Therefore, even though it has been claimed that collocations are accessed through associative links between its components (Hoey, 2005, 2012), these compositional links do not seem to depend on these features of the node word.

Overall, Study 2 (Chapter 5) confirms the pervasive influence of these linguistic factors on the acquisition of vocabulary, and indicates that interlexical factors might contribute to the difficulty of learning words in a stronger way than the intralexical factors.

9.1.3. Conceptualisation of vocabulary knowledge

Elaborating on the findings that vocabulary components seem to be mastered in a particular order, Studies 3 and 5 were designed to explore the contribution of each component to general vocabulary knowledge in an attempt to examine the nature of the overall vocabulary knowledge construct. In particular, the theoretically accepted multidimensionality of vocabulary knowledge was investigated in order to find empirical support for it. Based on the most widely accepted multidimensional description of vocabulary knowledge, Nation's (2013) framework, Study 3 in Chapter 6 tested a model in which the vocabulary components were specified as independent dimensions comprised of their recognition and recall masteries. The results determined that the so-called dimensions were very interlinked and could not be discerned as measurably different dimensions. This finding is important in that it indicates that the

multidimensionality of vocabulary knowledge seems to be only theoretical, and not a representation of the behaviour of vocabulary in the mental lexicon of Spanish learners of English. Rather, Study 3 (Chapter 6) provided empirical support for the unidimensional description of vocabulary knowledge, in which the individual word knowledge aspects contributed to vocabulary almost at the same level, but could not be considered separable dimensions, differing just in their degree of complication for learners. This finding demonstrates that when learning various word knowledge components, the type of knowledge being learnt interconnects with the previously known aspects of that word, improving their knowledge and expanding it (Meara, 1996). Thus, that interconnected knowledge cannot be separated in the mental lexicon, and therefore does not reflect independent dimensions in practice.

Yet, to demonstrate construct validity for the unidimensional vocabulary knowledge construct, this finding should be present in different language populations. Study 5 in Chapter 8 was designed for that purpose, and replicated the findings from Study 3 (Chapter 6) with Chinese learners of English. The results corroborated the lack of empirical support for a multidimensional understanding of vocabulary knowledge, in favour of a unidimensional model. The word knowledge components were found to be measuring the same construct, and with similar precision, across both language groups. The two studies together indicate that the alleged multidimensionality of vocabulary is not psychologically real, and that the best way of representing vocabulary knowledge in the mental lexicon of EFL learners is as a unidimensional construct where the various levels of knowledge difficulty interact with each other in the use of a language. This unidimensionality is particularly important for the theoretical discussion between size and depth of vocabulary knowledge, providing further evidence that they should not be treated as different constructs pragmatically.

This unidimensional result needs to be interpreted in light of its implications for testing and teaching in real language learning contexts. The similarly strong contribution of each word knowledge component to the unidimensional vocabulary knowledge construct highlights that, if the aim of measuring L2 vocabulary knowledge is to obtain accurate information about the degree to which a learner's vocabulary can help them use a language, then it is necessary to test multiple word knowledge components both in recognition and recall masteries. Although these formats cannot perfectly indicate the appropriate use of a word in reading or writing, they have been

found to be relatively good predictors of such skills (Cheng & Matthews, 2018; Li & Kirby, 2015; Milton, 2013; Qian, 2002). Consequently, assessing more word knowledge components at various levels of sensitivity would provide a better indication of what a learner could be able to do with a language.

Despite the empirical unidimensionality of vocabulary knowledge found in this thesis, it is also important to bear in mind, as Fitzpatrick and Clenton (2017, p. 2) state, that vocabulary "must in some manner be seen as multidimensional". In the particular case of assessing and teaching overall vocabulary, I agree that a multi-faceted view of vocabulary knowledge is possibly more useful. For example, it seems logical to consider the word knowledge components as different concepts in language teaching, as a learner may know the form–meaning link of a word but not much about its derivatives or collocations. In this sense, introducing the different components separately to students will be more practical for both learners and instructors. Nonetheless, this multidimensionality is not manifested in the learners' mental lexicon or the test/language behaviour, as learners cannot completely separate in their mind the different types of word knowledge in the use of language. This unidimensionality result is, thus, encouraging, as it shows that in order to use a word in real life appropriately, these types of knowledge need to interact with each other in the learners' mind (Meara, 1996; Perfetti, 2007).

To sum up, it has been claimed that the only way of obtaining informed answers about the acquisition of overall vocabulary knowledge is by adopting a model-driven approach to study the processes involved in vocabulary knowledge. My studies provide the first empirically-supported framework of vocabulary knowledge and acquisition, and hence can serve as the baseline model for the future study of vocabulary knowledge.

9.2. Directions for future research

The limitations of the studies presented in this thesis have been already addressed at the end of Chapters 4, 5, 6, 7 and 8. Therefore, this section will cover more general matters that can serve as suggestions for future studies. In my view, there are three main lines of enquiry which directly derive from the findings in this thesis and are worth pursuing in future research.

First of all, the analysis of vocabulary knowledge in this thesis is based only on cross-sectional examination of word knowledge components. Longitudinal data is needed to better understand how the different vocabulary components evolve individually and in relation to each other. Previous longitudinal research has shown that the incremental development of vocabulary aspects does not always follow a forward trend, but also regresses (Fitzpatrick, 2012). Yet, this finding is based on a very limited amount of informants, and data of various components was collected with only one test. Future research could examine the knowledge of multiple word knowledge components with various tests at multiple points in time to investigate the evolution of each component independently, as well as relative to the rest of the components. This examination would inform of how the order of acquisition of vocabulary components compares longitudinally and cross-sectionally. I would speculate that the order of individual components may vary, particularly with regard to multiple meanings and derivatives, but that the recognition-recall continuum would remain. This is still an empirical question to be answered, though.

The second line of research relates to overall vocabulary acquisition under different learning conditions. The findings in this thesis are based on the acquisition of vocabulary in general learning situations, where students from a wide range of learning contexts and environments gathered together. Research has shown that specific learning conditions and the type of learning tasks employed can make a difference in the knowledge of components (e.g., Dóczi & Kormos, 2016; Laufer & Rozovski-Roitblat, 2015; Webb, 2009). Hence, it is likely that the order of acquisition of word knowledge components found in this thesis varies when a specific treatment condition is applied. Future research should conduct diverse deliberate and incidental intervention studies in which the knowledge of various components is examined after a particular treatment in order to determine whether and how the specific vocabulary instruction vary the order of acquisition of components and their strength of relationship. One example would be to explore the effect that productive activities have on the recognition-recall scale. Previous research has found that these types of activities improve scores on the productive knowledge of components (e.g., Webb, 2005), but it remains unknown how they would affect the implicational scale found in this thesis. Along with the effect of treatment on the acquisition of vocabulary components, it is also possible that the development of vocabulary aspects in naturalistic settings differs. Thus, research is also

needed on the influence of a natural learning environment on the order and relationships between word knowledge components in order to better understand whether the results in this thesis are indicative of a natural order of acquisition.

The final avenue of future research concerns the generalisability of the findings to vocabulary knowledge as a universal phenomenon. The data analysed in this thesis derives only from learners of L2 English, limiting the findings to this language. Yet, a theory of vocabulary acquisition must be applicable to all languages. It is plausible that languages with similar lexical systems to English follow comparable routes in the acquisition of overall vocabulary, with more pronounced differences being evidenced in languages whose lexical systems differ considerably from that of English. In addition, some vocabulary aspects such as knowledge of derivatives or collocations could behave differently in languages where the morphological system is more complex (e.g., Turkish) or where word formation is not primarily morphemic and more-or-less independent concepts combine to create new meanings (e.g., Chinese). Consequently, the analysis of the order of acquisition of word knowledge components as well as the conceptualisation of the vocabulary knowledge construct in languages other than English would add to our knowledge of lexical acquisition and to the development of a general vocabulary learning theory.

9.3. Concluding remarks

While the multiple components comprising vocabulary knowledge have traditionally been claimed to be interrelated and constitute a multidimensional construct, research had not provided sufficient evidence for these claims. This thesis addressed the issue of the relationships between vocabulary components in order to obtain a better understanding of the nature of vocabulary knowledge. It showed that the development from recognition mastery to recall mastery seems to be a key issue in vocabulary acquisition, and that the performance of the interrelated degrees of word knowledge behaves similarly in different populations of English learners, regardless of the cognateness status of the learners' mother tongue. Although these results have highlighted some points of departure for interpreting the relationships between and learning of word knowledge components, there is a great deal more left to explore. Hence, these findings should be viewed as initial steps towards better understanding the

nature of general vocabulary knowledge and, as a result, developing a theory of its acquisition.

References

Abdi, H. (2010). Guttman scaling. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 559–561).

Abreu, J. L. (2012). Constructs, Variables, Dimensions, Indicators & Consistency. *Daena: International Journal of Good Conscience*, *7*(3), 123–130.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: John Wiley & Sons.

Agustín Llach, M., & Moreno Espinosa, S. (2014). Knowledge of form and meaning. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 60–72). Basingstoke: Palgrave Macmillan.

Aitchison, J. (2003). *Words in the mind.* (3ª). Oxford: Blackwell.

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. *English Lexicography in Japan*, 108–119.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409.

Alexander, R. (1982). What's in a four-letter word? Word meaning in English and second language learning. *Die Neueren Sprachen*, *81*(2), 219–224.

Andersen, R. W. (1978). An implicational model for second language research. *Language Learning*, *28*(2), 221–282.

Anderson, J. I. (1978). Order of difficulty in adult second language acquisition. In W. C. Ritchie (Ed.), *Second language acquisition research: Issues and implications* (pp. 91–108). New York: Academic Press, Inc.

Anderson, R., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (pp. 77–117). Newark, DE: International Reading Association.

Arbuckle, J. L. (2016). *Amos, Version 24.0.* [Computer Program]. Chicago: IBM SPSS.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20*(1), 50–57.

Awang, Z. (2012). *A handbook on SEM: Structural equation modelling using AMOS graphics*. Penerbit: Universiti Teknologi Mara.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Babayiğit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): A multigroup structural analysis. *Reading and Writing, 28*(4), 527–544.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74–94.

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

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101–114.

Bailey, C.-J. N. (1973). *Variation and linguistic theory*. Arlington, VA: Centre for Applied Linguistics.

Bailey, N., Madden, C., & Krashen, S. D. (1974). Is there a "natural sequence" in adult second language learning? *Language Learning, 24*(2), 235–243.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). New York, NY: Routledge.

Barclay, S. (2017). *The effect of word class and word length on the decay of lexical knowledge.* Presented at the American Association of Applied Linguistics, Portalnd, OR.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*(2), 323–363.

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language-multiple interpretations*. New York: Palgrave Macmillan.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Bolker, M. B. (2014). Package "lme4." *R Foundation for Statistical Computing, Vienna*, *12*.

Bauer, L., & Nation, I. S. P. (1993). Word Families. *International Journal of Lexicography*, *6*(3), 253–279.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). London: The Guilford Press.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations*. Amsterdam: John Benjamins.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588.

Blum-Kulka, S. (1981). Learning to use words: acquiring semantic competence in a second language. In M. Nahir (Ed.), *Hebrew Teaching and Applied Linguistic*. Washington, DC: University Press of America.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. A. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, *18*(1), 54–74.

Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Basingstoke: Palgrave Macmillan.

Bollen, K., & Long, J. (1993). Introduction. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 1–9). Newbury Park, CA: Sage Publications.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*(3), 461–483.

Bravo, M. A., Hiebert, E. H., & Pearson, P. D. (2005). Tapping the Linguistic Resources of Spanish–English Bilinguals. In R. Wanger, A. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary development and its implications for reading comprehension* (pp. 140–165). New York, NY: Guilford.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*(1), 1–22.

Brooks, R. T. (2014). Cronbach's Alpha below 0.7. Retrieved from https://www.researchgate.net/post/Cronbachs_alpha_below_07

Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research*, *15*(1), 83–97.

Brown, D. (2013). Types of words identified as unknown by L2 learners when reading. *System*, *41*(4), 1043–1055.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *37*(4), 13–16.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). New York: Routledge.

Campoy, G. (2008). The effect of word length in short-term memory: Is rehearsal necessary? *The Quarterly Journal of Experimental Psychology*, *61*(5), 724–734.

Carlisle, J. F. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly*, *45*(4), 464–487.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., … White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, *39*(2), 188–215.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

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

Chen, X., Ramirez, G., Luo, Y. C., Geva, E., & Ku, Y.-M. (2012). Comparing vocabulary development in Spanish-and Chinese-speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing*, *25*(8), 1991–2020.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Chondrogianni, V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism*, *1*(3), 318–345.

Chui, A. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching*, *16*, 1–23.

Churchill, E. (2008). A dynamic systems account of learning a word: From ecology to form relations. *Applied Linguistics*, *29*(3), 339–358.

Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2*

Vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis. (Vol. 2, pp. 79–108). Eurosla Monograph Series.

Cobb, T. (2015). *Web Vocab Profiler Compleat*. Retrieved from https://www.lextutor.ca/vp/comp/

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

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72–89.

Cowie, A. P. (1994). Phraseology. In R.E. Asher & J.M.Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (Vol. 6, pp. 3168–3171).

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

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149–161). Amsterdam: John Benjamins.

Cremer, M., Dingshoff, D., de Beer, M., & Schoonen, R. (2010). Do word associations assess word knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism, 15*(2), 187–204.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research, 36*(3), 206–217.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573–605.

Cummins, J. (1979). Cognitive/Academic Language Proficiency, Linguistic Interdependence, the Optimum Age Question and Some Other Matters. *Working Papers on Bilingualism, 19*, 198–203.

Cunningham, T. H., & Graham, C. R. (2000). Increasing native English vocabulary recognition through Spanish immersion: Cognate transfer from foreign to first language. *Journal of Educational Psychology, 92*(1), 37–49.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369–382.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*(8), 895–948.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: University Press Cambridge.

David, A., Myles, F., Rogers, V., & Rule, S. (2009). Lexical development in instructed L2 learners of French: is there a relationship with morphosyntactic development? In B. Richards, H. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 147–163). Basingstoke: Palgrave Macmillan.

Davidson, F. (1987). *Guttman scalogram analysis*. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present. Retrieved from http://corpus.byu.edu/coca/.

Davies, Mark, & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. London: Routledge.

De Groot, A. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*(3), 463–506.

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

De Groot, A. M., Borgwaldt, S., Bos, M., & Van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language*, *47*(1), 91–124.

De Groot, A., & Van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

DeCamp, D. (1971). Toward a generative analysis of a post-creole speech continuum. In D. Hymes (Ed.), *Pidginization and creolization of languages* (pp. 349–370). Cambridge, MA: Cambridge University Press.

Deshon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137–149.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). London: Sage publications.

Dickerson, L. J. H. (1975). Internal and external patterning of phonological variability in the speech of Japanese learners of English: Toward a theory of second-language acquisition. *Dissertation Abstracts International. A. The Humanities and Social Sciences*, *35*(11), 7285–7286.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology, 66*(5), 843–863.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(2), 119–143.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford: Oxford University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Dörnyei, Z., Ibrahim, Z., & Muir, C. (2015). "Directed Motivational Currents": Regulating complex dynamic systems through motivational surges. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 95–105). Bristol: Multilingual Matters.

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

Dulay, H. C., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*(1), 37–53.

Duñabeitia, J. A., Cholin, J., Corral, J., Perea, M., & Carreiras, M. (2010). SYLLABARIUM: An online application for deriving complete statistics for Basque and Spanish orthographic syllables. *Behavior Research Methods*, *42*(1), 118–125.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test (PPVT-IV)* (4th ed.). Circle Pines, MN: American Guidance Service.

Dunn-Rankin, P. (1983). *Scaling Methods*. Hillsdale, NJ: Lawrence Erlbaum.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169.

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

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

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

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

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology*, *46*(3), 533–558.

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. *Applied Linguistics*, *27*(4), 558–589.

Ellis, R. (1989). Are classroom and naturalistic acquisition the same? *Studies in Second Language Acquisition*, *11*(03), 305–328.

Ellis, R. (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics*, *18*(1), 4–22.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509–529.

Farghal, M., & Obiedat, H. (1995). Collocations: A neglected variable in EFL. *IRAL-International Review of Applied Linguistics in Language Teaching*, *33*(4), 315–332.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: Sage.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: IAP.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford: Oxford University Press.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EuroSLA Yearbook*, *6*(1), 121–145.

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. (2017). Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge. *TESOL Quarterly, Online Fir*, 1–24. https://doi.org/10.1002/tesq.356

Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2. *Studies in Second Language Acquisition*, *33*(03), 373–398.

Fitzpatrick, T., & Milton, J. (2014). Conclusion: reconstructing vocabulary knowledge. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 173–177). Basingstoke: Palgrave Macmillan.

Fitzpatrick, T., & Munby, I. (2014). Knowledge of word associations. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 92–105). Basingstoke: Palgrave Macmillan.

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

Fox, J., & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, *32*(1), 1–24.

Fracasso, L. E., Bangs, K., & Binder, K. S. (2016). The contributions of phonological and morphological awareness to literacy skills in the adult basic education population. *Journal of Learning Disabilities*, *49*(2), 140–151.

Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal*, *91*(3), 372–394.

García, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*, *26*(1), 371–392.

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

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-Acquisition, Word Frequency, and Neighborhood Density Effects on Spoken Word Recognition by Children and Adults. *Journal of Memory and Language*, *45*(3), 468–492.

Garnier, M. (2016). *English phrasal verbs: Usage, knowledge, acquisition* (Unpublished PhD thesis). The University of Nottingham.

Gibson, M. L. (1998). *Dialect contact in Tunisian Arabic: Sociolinguistic and structural aspects* (Unpublished PhD thesis). The University of Reading.

Gitsaki, C. (1996). *The development of ESL collocational knowledge* (Unpublished PhD thesis). University of Queensland.

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: the effects of frequency and amount of exposure. *ITL-Internationa*

l Journal of Applied Linguistics, *166*(1), 94–126.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, Analysis, and Applications*, 79–100.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 191–208). Amsterdam: John Benjamins.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150.

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 153–170). New York: Palgrave Macmillan.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective–challenges and potential solutions. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 Vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis.* (Vol. 2, pp. 11–28). Eurosla Monograph Series.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278–306.

Gyllstad, H., & Wolter, B. (2015). Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter? *Language Learning*.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, *10*(2), 221–240.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). New York: Pearson.

Hair, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). London: Sage Publications.

Hall, C. J. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *International Review of Applied Lingustics in Language Teaching*, *40*(2), 69–88.

Halliday, M. A. (1994). *An Introduction to Functional grammar*. London: Edward Arnold.

Hancin-Bhatt, B., & Nagy, W. E. (1994). Lexical transfer and second language morphological development. *Applied Psycholinguistics*, *15*(3), 289–310.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House Publishers.

Hayashi, Y., & Murphy, V. (2011). An investigation of morphological awareness in Japanese learners of English. *Language Learning Journal*, *39*(1), 105–120.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(02), 303–317.

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language* (pp. 22–66). New York: Palgrave Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development – A progress report. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (Vol. 2, pp. 29–56). Amsterdam: John Benjamins.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hoey, M. (2012). Lexical priming. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Hornby, A. S. (1954). *Guide to patterns and usage in English*. Oxford: Oxford University Press.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring Second Language Vocabulary through Reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Hoyle, R. H., & Panter, A. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huchinson, T. (2015). Low-confidence responses on the vocabulary size test. *Vocabulary Learning and Instruction*, *4*(2), 49–51.

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

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

IBM Corp. (2016). *IBM SPSS Statistics for Windows, Version 24.0*. Armonk, NY: IBM Corp.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*(3), 250–276.

Irujo, S. (1993). Steering clear: Avoidance in the production of idioms. *Iral-International Review of Applied Linguistics in Language Teaching*, *31*(3), 205–220.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

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

Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*(4), 617–638.

Jiménez, R. T., García, G. E., & Pearson, P. D. (1996). The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly*, *31*(1), 90–112.

Johnson, D. E. (2013). Descriptive statistics. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 288–315). Cambridge: Cambridge University Press.

Joyce, P. (2015). L2 vocabulary learning and testing: The use of L1 translation versus L2 definition. *The Language Learning Journal*, 1–12.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*(9), 1131–1142.

Kenny, D. A. (2016). Multiple Latent Variable Models: Confirmatory Factor Analysis. Retrieved June 12, 2016, from http://davidakenny.net/cm/mfactor.htm

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507.

Khodadady, E., & Khaghaninizhad, M. S. (2012). Acquisition of French polysemous vocabularies: Schema-based instruction versus translation based instruction. *Porta Linguarum*, *17*, 29–46.

Kieffer, M. J., & Lesaux, N. K. (2008). The role of derivational morphology in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing*, *21*(8), 783–804.

Kieffer, M. J., & Lesaux, N. K. (2012a). Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade. *Reading and Writing*, *25*(2), 347–373.

Kieffer, M. J., & Lesaux, N. K. (2012b). Development of morphological awareness and vocabulary knowledge in Spanish-speaking language minority learners: A parallel process latent growth curve model. *Applied Psycholinguistics*, *33*(01), 23–54.

Kieffer, M. J., & Lesaux, N. K. (2012c). Direct and Indirect Roles of Morphological Awareness in the English Reading Comprehension of Native English, Spanish, Filipino, and Vietnamese Speakers. *Language Learning*, *62*(4), 1170–1204.

Kim, S. H., & Kim, J. H. (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly*, *46*(4), 831–841.

Kim, Y. (2008). The contribution of collaborative and individual tasks to the acquisition of L2 vocabulary. *The Modern Language Journal*, *92*(1), 114–130.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: The Guilford Press.

Koda, K. (2008). Impacts of prior literacy experience on second language learning to read. In K. Koda & M. Zehler (Eds.), *Learning to read across languages: Cross-linguistic relationships in first-and second-language literacy development* (pp. 68–96). New York: Taylor & Francis.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2016). MVN: An R package for assessing multivariate normality. *The R Journal*, *6*(2), 151–162.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848–870.

Kremmel, B., & Schmitt, N. (2016). Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us About Learners' Ability to Employ Words? *Language Assessment Quarterly*, *13*(4), 377–392.

Kremmel, B., & Schmitt, N. (2017). Vocabulary Levels Test. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0499

Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, *18*(2), 137–171.

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, *41*(3), 161–180.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models. R Package Version 2.0-33* [Retrieved from https://CRAN.R-project.org/package=lmerTest].

Larsen-Freeman, D. E. L. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, 409–419.

Laufer, B. (1997). What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, *19*(2), 255–271.

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world a response to Meara (2005). *Applied Linguistics*, *26*(4), 582–588.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716.

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

Laufer, B., & Nation, P. (2011). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). London: Routledge.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.

Laufer, B., & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? In A. N. S. Foster-Cohen (Ed.), *Eurosla Yearbook* (Vol. 1, pp. 7–28). Amsterdam: John Benjamins.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391.

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

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*(4), 391–411.

Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research*, *19*(6), 687–711.

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

Laura Vilkaitė. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1632–1642.

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, *31*(4), 537–561.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, *40*(2), 295–320.

Lee, S., & Pulido, D. (2017). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research*, *21*(1), 118–135.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Routledge.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. *EUROSLA Monographs Series*, *2*, 127–148.

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

Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*(5), 611–634.

Lo, Y., Anderson, A. L., & Bunch-Crump, K. (2017). Building vocabulary of English learners with reading disabilities through computer-assisted morphology instruction. *Intervention in School and Clinic*, *52*(3), 133–140.

Lubliner, S., & Hiebert, E. H. (2011). An Analysis of English–Spanish Cognates as a Source of General Academic Language. *Bilingual Research Journal*, *34*(1), 76–93.

Ma, Q. (2015). A process-focused learning model for L2 vocabulary acquisition: Construction, implementation and validation. *ITL-International Journal of Applied Linguistics*, *166*(1), 127–162.

Macis, M., & Schmitt, N. (2017a). Not just "Small Potatoes": Knowledge of the idiomatic meanings of collocations. *Language Teaching Research*, *21*(3), 321–340.

Macis, M., & Schmitt, N. (2017b). The figurative and polysemous nature of collocations and their place in ELT. *Elt Journal*, *71*(1), 50–59.

Malabonga, V., Kenyon, D. M., Carlo, M., August, D., & Louguit, M. (2008). Development of a cognate awareness measure for Spanish-speaking English language learners. *Language Testing*, *25*(4), 495–519.

Manning, C. (2007). Logistic regression (with R). Retrieved from http://nlp. stanford. edu.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 519–530.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, Online Fir, 1-71. https://doi.org/10.1111/lang.12286

McBride-Chang, C., Tardif, T., Cho, J.-R., Shu, H., Fletcher, P., Stokes, S. F., … Leung, K. (2008). What's in a word? Morphological awareness and vocabulary knowledge in three languages. Applied Psycholinguistics, 29(03), 437–462.

McCutchen, D., Green, L., & Abbott, R. D. (2008). Children's morphological knowledge: Links to literacy. Reading Psychology, 29(4), 289–314.

McCutchen, D., & Stull, S. (2015). Morphological awareness and children's writing: accuracy, error, and invention. Reading and Writing, 28(2), 271–289.

McDonald, K., & Asaba, M. (2016). "'I Don't Know"'Use and Guessing on the Bilingual Japanese Vocabulary Size Test: Clarifications and Limitations. Vocabulary Learning and Instruction, 5(1), 34–39.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. Psychological Methods, 7(1), 64–82.

McGee, I. (2008). Word frequency estimates revisited—A response to Alderson (2007). Applied Linguistics, 29(3), 509–514.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. Corpus Linguistics and Linguistic Theory, 5(1), 79–103.

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. Shiken, 19(2), 1–11.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology, 93(3), 568.

Meara, P. (1983). Vocabulary in a Second Language, Vol. 1. Specialised Bibliography 3. London: CILT.

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). Tampere, Finland: Publications de l'Association Finlandaise de Linguistique Apllique'e.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press.

Meara, P. (2009). Connected words: Word associations and second language vocabulary acquisition. Amsterdam: John Benjamins Publishing.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. System, 28(1), 19–30.

Meara, P., & Jones, G. (1990). Eurocentres vocabulary size test 10KA. Eurocentres Learning Service.

Meara, P., & Milton, J. (2003). X_Lex: the Swansea vocabulary levels test. Newbury: Express Publishing.

Meara, P., & Miralpeix, I. (2006). Y_Lex: The Swansea advanced vocabulary levels test. v2. 05. Swansea: Lognostics.

Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), Angles on the English speaking world 4 (pp. 85–96). Copenhagen: Museum Tusculanum Press.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, acquisition and pedagogy (Vol. 33, pp. 84–102). Cambridge: Cambridge University Press.

Méndez Pérez, A., Peña, E. D., & Bedore, L. M. (2010). Cognates facilitate word recognition in young Spanish-English bilinguals' test performance. Early Childhood Services (San Diego), 4(1), 55–67.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156–165.

Milton, J. (2009). Measuring second language vocabulary acquisition. Bristol: Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C Bardel, C. Lindqvist, & B. Laufer (Eds.), L2 vocabulary acquisition, knowledge and use (pp. 57–78). European Second Langauge Association.

Milton, J., & Fitzpatrick, T. (2014). Introduction: Deconstructing vocabulary knowledge. In J. Milton & T. Fitzpatrick (Eds.), Dimensions of vocabulary knowledge (pp. 1–12). Basingstoke: Palgrave Macmillan.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? Canadian Modern Language Review, 63(1), 127–147.

Milton, J., & Riordan, O. (2006). Level and script effects in the phonological and orthographic vocabulary size of Arabic and Farsi speakers. In P. Davidson, C.

Coombe, D. Lloyd, & D. Palfreyman (Eds.), Teaching and learning vocabulary in another language (pp. 122–133). UAE: TESOL Arabia.

Miralpeix, I. (2006). Age and vocabulary acquisition in English as a Foreign Language (EFL). In C. Muñoz (Ed.), Age and the rate of foreign language learning (Vol. 19, pp. 89–106). Clevedon: Multilingual Matters.

Mizrahi, E. (2016). Lexical competence of highly advanced L2 users and its relation to leaning experience and learner variables (Unpublished PhD thesis). University of Haifa.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. System, 28(2), 291–304.

Molenaar, I. W., & Sijtsma, K. (2000). User's manual MSP5 for Windows [software manual]. Groningen, The Netherlands: iecProGAMMA.

Moll, L. C., Amanti, C., Neff, D., & González, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. Theory into Practice, 31(2), 132–141.

Molnár, T. (2010). Cognate recognition and L3 vocabulary acquisition. Acta Universitatis Sapientiae, Philologica, 2(2), 337–349.

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. Language Learning & Technology, 18, 118–141.

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 371–383). New York, NY: Routledge.

Nagy, W. E., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. Journal of Educational Psychology, 9(1), 134–147.

Nagy, W. E., Diakidoy, I.-A. N., & Anderson, R. C. (1993). The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. Journal of Literacy Research, 25(2), 155–170.

Nagy, W. E., García, G. E., Durgunoğlu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. Journal of Reading Behavior, 25(3), 241–259.

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

Nagy, W. E., McClure, E. F., & Mir, M. (1995). Linguistic transfer and the use of context by Spanish-English bilinguals. Champagne, IL: Centre for the Study of Reading.

Nash, R. (1997). NTC's dictionary of Spanish cognates. Chicago, IL: NTC Publishing Group.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? Canadian Modern Language Review, 63(1), 59–82.

Nation, P. (1990). Teaching and learning vocabulary. New York: Newbury House.

Nation, P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge: Cambridge University Press.

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

Nation, P., & Webb, S. A. (2011). Researching and analyzing vocabulary. Boston, MA: Heinle.

Nation, P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. TESOL Journal, 9(2), 6–10.

Nation, P. (2007). The four strands. International Journal of Innovation in Language Learning and Teaching, 1(1), 2–13.

Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.

Nemati, A. (2010). Active and passive vocabulary knowledge: The effect of years of instruction. The Asian EFL Journal Quarterly, 12(1), 30–46.

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

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

Newman, M. (2010). Focusing, implicational scaling, and the dialect status of New York Latino English. Journal of Sociolinguistics, 14(2), 207–239.

Nguyen, T. M. H., & Webb, S. A. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. Language Teaching Research, 21(3), 298–320.

Nissen, H. B., & Henriksen, B. (2006). Word class influence on word association test results1. International Journal of Applied Linguistics, 16(3), 389–408.

Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. ARELE (Annual Review of English Language Education in Japan), 13, 71–80.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. English for Specific Purposes, 18(2), 161–175.

Otwinowska-Kasztelanic, A. (2009). Raising awareness of cognate vocabulary as a strategy in teaching English to Polish adults. International Journal of Innovation in Language Learning and Teaching, 3(2), 131–147.

Palmer, H. E. (1917). The scientific study and teaching of language. London: Harrap.

Palmer, H. E. (1921). The principles of language study. London: Harrap.

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. TESL Canada Journal, 11(1), 09-29.

Pellicer-Sánchez, A. (2016). Incidental L2 Vocabulary Acquisition from and while Reading: An Eye-Tracking Study. Studies in Second Language Acquisition, 38(1), 97–130.

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. Language Teaching Research, 21(3), 381–402. https://doi.org/10.1177/1362168815618428

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: do things fall apart? Reading in a Foreign Language, 22(1), 31.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. Language Testing, 29(4), 489–509.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In E. Gough, P. B. L. C. & Treiman, R. (Eds.), Reading Acquisition (pp. 145–174). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. Journal of Experimental Psychology: Learning Memory and Cognition, 31, 1281–1292.

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

Pienemann, M. (1998). Language processing and second language development: Processability theory (Vol. 15). Amsterdam: John Benjamins Publishing.

Pienemann, M. (2005). Cross-linguistic aspects of Processability Theory (Vol. 30). John Benjamins Publishing.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. Reading in a Foreign Language, 18(1), 1–28.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. Language Learning, 64(4), 878–912.

Politzer, R. L. (1976). The implicational relation paradigm in language acquisition. In Juilland (Ed.), Linguistic studies offered to Joseph Greenberg (pp. 123–135). Saratoga, CA: Anma Libri.

Porte, G. (Ed.). (2012). Replication research in Applied Linguistics. Cambridge: Cambridge University Press.

Porte, G., & Richards, K. (2012). Focus article: Replication in second language writing research. Journal of Second Language Writing, 21(3), 284–293.

Proctor, C. P., & Mo, E. (2009). The Relationship Between Cognate Awareness and English Comprehension Among Spanish—English Bilingual Fourth Grade Students. TESOL Quarterly, 43(1), 126–136.

Purpura, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. Language Learning, 47(2), 289–325.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. Canadian Modern Language Review, 56(2), 282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. Language Learning, 52(3), 513–536.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. Language Testing, 21(1), 28–52.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. Journal of Memory and Language, 59(4), 413–425.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Ramírez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-speaking English language learners: The facilitating effect of cognate knowledge. Topics in Language Disorders, 33(1), 73–92.

Ramzan, S., Zahid, F. M., & Ramzan, S. (2013). Evaluating multivariate normality: A graphical approach. Middle East Journal of Scientific Research, 13(2), 254–263.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. Language Testing, 10(3), 355–371.

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

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), Validation in language assessment (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum.

Read, J. (2000). Assessing vocabulary. Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In P. Bogaards & B. Laufer (Eds.), Vocabulary in a second language: Selection, acquisition, and testing (pp. 209–227). Amsterdam: John Benjamins.

Real Academia. (2015). Corpus del Español del Siglo XXI. Real Academia Española. Banco de datos (CORPES XXI). Retrieved from http://web.frl.es/CORPES/view/inicioExterno.view;jsessionid=C2EC53A5A6C877037CE1FE6EF0B1E9C3

Redwood, S., & Schmitt, N. (2011). In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), A Taste for Corpora. Amsterdam: John Benjamins.

Reynolds, B. L. (2015). The effects of word form variation and frequency on second language incidental vocabulary acquisition through reading. Applied Linguistics Review, 6(4), 467–497.

Richards, J. C. (1976). The role of vocabulary teaching. TESOL Quarterly, 10, 77–89.

Richards, J. C. (1980). Second language acquisition: Error analysis. Annual Review of Applied Linguistics, 1, 91–107.

Rickford, J. R. (1975). Carrying the new wave into syntax: The case of Black English BIN. In R. W. Fasold & R. W. Shuy (Eds.), Analyzing variation in English (pp. 162–183). Washington, DC: Georgetown University Press.

Rickford, J. R. (2002). Implicational scales. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), The handbook of language variation and change (pp. 142–167). Malden, Massachusetts: Blackwell.

Rodríguez, T. A. (2001). From the known to the unknown: Using cognates to teach English to Spanish-speaking literates. The Reading Teacher, 54(8), 744–746.

Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? Language Teaching Research, 19(1), 9–27.

Rogers, V. (2009). Syntactic development in the second language acquisition of French by instructed English learners (Unpublished PhD thesis). University of Newcastle Upon Tyne.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1–36.

Rosselli, M., Ardila, A., Jurado, M. B., & Salvatierra, J. L. (2012). Cognate facilitation effect in balanced and non-balanced Spanish–English bilinguals using the Boston Naming Test. International Journal of Bilingualism, 18(6), 649–662.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. Personality and Social Psychology Bulletin, 28(12), 1629–1646.

Saigh, K. (2015). Derivative knowledge and the development of an affix list for pedagogy (Unpublished PhD thesis). University of Nottingham.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. Second Language Research, 27(3), 343–360.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods of Psychological Research Online, 8(2), 23–74.

Schmitt, N. (1997). Using a word knowledge framework to research vocabulary (Unpublished PhD thesis). The University of Nottingham.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. Language Learning, 48(2), 281–317.

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

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

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

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

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. Second Language Research, 15(4), 389–411.

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

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. Word associations and verbal suffixes. Studies in Second Language Acquisition, 19(01), 17–36.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Language Teaching, 47(04), 484–503.

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.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? TESOL Quarterly, 145–171.

Schoonen, R., Gelderen, A. van, Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. Language Learning, 53(1), 165–202.

Scott, J. A., Flinspach, S. L., Vevea, J. L., & Castaneda, R. (2017). Vocabulary knowledge as a multidimensional concept: A six factor model. Poster presentation. Society for the Scientific Study of Reading Conference.

Scott, J. A., Miller, T. F., & Flinspach, S. L. (2012). Developing Word Consciousness: Lessons from highly diverse fourth-grade classrooms. In E. J. Kame'enui & J. F. Baumann (Eds.), Vocabulary instruction: Research to practice (2nd ed., pp. 169–188). London: The Guilford Press.

Segalowitz, N. S., & Gatbonton, E. (1977). Studies of the nonfluent bilingual. Bilingualism: Psychological and Social Implications, 77–90.

Segalowitz, N. S., & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll (Ed.), Handbook of Bilingualism: Psycholinguistic Approaches (pp. 371–388). Oxford: Oxford University Press.

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

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. Applied Psycholinguistics, 19(01), 53–67.

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. JABAET Journal, 4, 69–80.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. Language Testing, 24(1), 99–128.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. Psychometrika, 52(1), 79–97.

Sijtsma, K., & Molenaar, I. W. (2002). Introduction to nonparametric item response theory (Vol. 5). London: Sage Publications.

Simpson Baird, A., Palacios, N., & Kibler, A. (2016). The Cognate and False Cognate Knowledge of Young Emergent Bilinguals. Language Learning, 66(2), 448–470.

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

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Siyambalapitiya, S., Chenery, H. J., & Copland, D. A. (2009). Bilingualism and aging: Reversal of the cognate advantage in older bilingual adults. Applied Psycholinguistics, 30(3), 531–554.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. Canadian Modern Language Review, 64(3), 429–458.

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. Language Learning, 65(3), 533–562.

Spencer, M., Muse, A., Wagner, R. K., Foorman, B., Petscher, Y., Schatschneider, C., … Bishop, M. D. (2015). Examining the underlying dimensions of morphological awareness and vocabulary knowledge. Reading and Writing, 28(7), 959–988.

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

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

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. Quality & Quantity, 43(4), 599–616.

Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. Vocabulary Learning and Instruction, 1(1), 53–59.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? Language Assessment Quarterly, 11(3), 271–282.

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT

method in empirical research for applied health researchers. BMC Medical Research Methodology, 12(1), 74.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 471–483). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. International Journal of Applied Linguistics, 26(2), 245–265.

Takala, S. (1985). Estimating students' vocabulary sizes in foreign language teaching. In V. Kohonen, H. van Essen, & C. Klein-Braley (Eds.), Practice and problems in language testing. (pp. 157–165). Tampere, Finland: AfinLA.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. Scientific Studies of Reading, 10(4), 381–398.

Thornbury, S. (2002). How to Teach Vocabulary. Harlow: Longman.

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

Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. Applied Linguistics, 1–27.

Tseng, W., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. Applied Linguistics, 27(1), 78–102.

Tseng, W., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. Language Learning, 58(2), 357–400.

Ullman, J. (2013). Structural equation modeling. In B. Tabachnick & L. Fidell (Eds.), Using multivariate statistics (6th ed., pp. 681–778). Boston, MA: Pearson.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. Bilingualism: Language and Cognition, 1(3), 193–211.

Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. Political Analysis, 11(2), 139–163.

Van Schuur, W. H. (2011). Ordinal Item Response Theory: Mokken scale analysis. Thousand Oaks, CA: Sage.

Van Zeeland, H. (2014). Second language vocabulary knowledge in and from listening (Unpublished PhD thesis). University of Nottingham, UK.

Van Zeeland, H., & Schmitt, N. (2013a). Incidental vocabulary acquisition through L2 listening: A dimensions approach. System, 41(3), 609–624.

Van Zeeland, H., & Schmitt, N. (2013b). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? Applied Linguistics, 34(4), 457–479.

Vandenberg, R. J. (2006). Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea? Organizational Research Methods, 9(2), 194–201.

VanPatten, B., Williams, J., & Rott, S. (2004). Form-meaning connections in second language acquisition. In J. W. B. VanPatten & M. O. S. Rott (Eds.), Form-meaning connections in second language acquisition (pp. 1–26). Mahwah: Erlbaum Press.

Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. Applied Linguistics, 14(4), 344–363.

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

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. Language Learning, 53(3), 547–586.

Vidal, K. (2003). Academic listening: A source of vocabulary acquisition? Applied Linguistics, 24(1), 56–89.

Vidal, K. (2011). A Comparison of the Effects of Reading and Listening on Incidental Vocabulary Acquisition. Language Learning, 61(1), 219–258.

Vilkaitė, L., & Schmitt, N. (2017). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? Applied Linguistics. https://doi.org/https://doi.org/10.1093/applin/amx030

Wallace, G., & Hammill, D. D. (2002). CREVT3: Comprehensive Receptive and Expressive Vocabulary Test (2nd ed.). Austin, TX: Pro-Ed.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. System, 37(3), 461–469.

Webb, S. A. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. Studies in Second Language Acquisition, 27(01), 33–52.

Webb, S. A. (2007a). The effects of repetition on vocabulary knowledge. Applied Linguistics, 28(1), 46–65.

Webb, S. A. (2007b). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. Language Teaching Research, 11(1), 63–81.

Webb, S. A. (2008). Receptive and productive vocabulary sizes of L2 learners. Studies in Second Language Acquisition, 30(01), 79–95.

Webb, S. A. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. RELC Journal, 40(3), 360–376.

Webb, S. A., & Nation, P. (2017). How vocabulary is learned. Oxford: Oxford University Press.

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

Webb, S. A., & Rodgers, M. P. (2009). The lexical coverage of movies. Applied Linguistics, 30(3), 407–427.

Webb, S. A., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. ITL-International Journal of Applied Linguistics, 168(1), 33–69.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. Canadian Modern Language Review, 53(1), 13–40.

West, M. (1953). A General Service List of English Words. London: Longman.

Wilks, C. (2014). Knowledge of constraints on use. In J. Milton & T. Fitzpatrick (Eds.), Dimensions of vocabulary knowledge (pp. 140–153). Basingstoke: Palgrave Macmillan.

Wilks, C., & Meara, P. (2007). Graph theory and word association networks. In H Daller, J. Milton, & J. Treffers-Daller (Eds.), Modelling and assessing vocabulary knowledge (pp. 167–181). Cambridge: Cambridge University Press.

Williams, J. N., & Cheung, A. (2011). Using priming to explore early word learning. In M. Trofimovich, P. K. (Ed.), Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics (pp. 73–104). Amsterdam: John Benjamins Publishing.

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. The Language Learning Journal, 40(1), 125–137.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. In University of California, Merced. Cognitive and Information Sciences.

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope? System, 30(3), 315–329.

Wolter, B. (2009). Meaning-last vocabulary acquisition and collocational productivity. In T. Fitzpatrick & A. Barfield (Eds.), Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara (pp. 111–127). Bristol: Multilingual Matters.

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

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. Studies in Second Language Acquisition, 35(3), 451–482.

Wolter, B., & Yamashita, J. (2017). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? Studies in Second Language Acquisition, 1–22.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.

Wray, A. (2008). Formulaic Language: Pushing the boundaries. Oxford: Oxford University Press.

Yuksel, D., & Tanriverdi, B. (2009). Effects of watching captioned movie clip on vocabulary development of EFL learners. TOJET: The Turkish Online Journal of Educational Technology, 8(2), 1–7.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. Canadian Modern Language Review, 57(4), 541–572.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. The Modern Language Journal, 96(4), 558–575.

Zhang, D. (2016). Morphology in Malay–English biliteracy acquisition: an intervention study. International Journal of Bilingual Education and Bilingualism, 19(5), 546–562.

Zhang, D., & Koda, K. (2011). Contribution of morphological awareness and lexical inferencing ability to L2 vocabulary knowledge and reading comprehension among advanced EFL learners: testing direct and indirect effects. Reading and Writing, 25(5), 1195–1216.

Zhang, D., & Yang, X. (2016). Chinese L2 Learners' Depth of Vocabulary Knowledge and Its Role in Reading Comprehension. Foreign Language Annals, 49(4), 699–715.

Zhang, H. (2016). Does Morphology Play an Important Role in L2 Chinese Vocabulary Acquisition? Foreign Language Annals, 49(2), 384–402.

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. TESOL Quarterly, 47(4), 790–811.

Zoccolotti, P., De Luca, M., Di Filippo, G., Judica, A., & Martelli, M. (2009). Reading development in an orthographically regular language: Effects of length, frequency, lexicality and global processing ability. Reading and Writing, 22(9), 1053–1079.

Appendices

Appendix 1. Tentative order of acquisition of multicomponent studies.

The following table is an attempt to extract an acquisition order from the results in multicomponent studies. It summarises the author(s)' conclusions in these studies and suggests what appears to be the acquisition order of word knowledge components based on the results.

Study	Degree of knowledge and Tentative order of components
Schmitt (1998)	Author's conclusion on the degree of knowledge: Spelling Recall was the best known aspect, with steady gains. Recall knowledge of multiple meanings remained stable, with few gains. Associations recall gains at same rate as meaning gains. Grammatical information (word class + derivatives) recall known better than associations, but gains at the same rate as associations. **** Tentative order of components according to knowledge demonstrated in the final test administration (mean % scores) for the 3 participants: 1 – *Spelling (97%) > Grammatical Information (73%) > Associations (70%) > Meanings (56%)* 2 – *Spelling (91%) > Associations (58%) > Grammatical Information (55%) > Meanings (36%)* 3 – *Spelling (100%) > Grammatical information (84%) > Associations (73%) > Meanings (44%)* General tentative order for all components: *Spelling > Grammatical Information ≈ Associations > Meanings*

Chui (2006)	Author's conclusion on the degree of knowledge: Word class recognition was the best known, followed by meaning recognition. Limited derivative productive and collocation recognition knowledge. The only productive aspect measured (derivatives) was the worst known. **** Tentative order based on the mean % scores: *Word class recognition (88%) > Meaning recognition (65%) > Collocation recognition (57%) > Derivatives production (52%)*
Schmitt & Meara (1997)	Author's conclusion on the degree of knowledge: Poor knowledge and gains in suffixes, specially derivatives (they knew only 15% of productive derivatives, and 40% of receptive derivatives). Not very good knowledge of associations, with very low gains. Receptive knowledge was generally better than productive in each component, but with no stable growth patterns between R and P gains. **** Tentative order according to knowledge demonstrated in the final test administration (mean % scores) *Receptive Suffixes (66%) > Productive Suffixes (47%) > Receptive Associations (43%) > Productive Associations (25%)*
Pigada & Schmitt (2006)	Author's conclusion on the degree of knowledge: Spelling recall experienced the greatest gains even from few exposures. Meaning recall and grammatical knowledge (recognition for nouns, recall for verbs) increased, but at a lower rate. ****

	Tentative order according to knowledge demonstrated in the final test administration (mean % scores) *Spelling Recall (60%) > Grammatical Knowledge recognition (43%) > Meaning Recall (24%) > Grammatical Knowledge Recall (21%).*
Webb (2005)	Author's conclusion on the degree of knowledge: Orthography experienced the highest gains, both receptively and productively. Grammar tends to get high scores. Associations is generally around the middle point. Syntax and form–meaning link tend to get lower scores. **** Tentative orders according to combined receptive and productive mean % scores for each treatment ***1st experiment:*** writing vs reading with equal time in both tasks Writing: *Grammar (87%) > Orthography (84%) > Meaning and form (82%) > Associations (81%) > Syntax (71%)* Reading: *Orthography (93%) = Grammar (93%) = Meaning and Form (93%) > Association (89%) > Syntax (80%)* ***2nd experiment:*** writing vs reading with more time in the writing task than reading Writing: *Grammar (54%) > Orthography (50%) > Associations (47%) > Syntax (42%) > Meaning and form (41%)* Reading: *Grammar (43%) > Orthography (40%) > Association (34%) > Syntax (33%) > Meaning and Form (24%)*

The above orders show the receptive and productive scores combined. When we looked at the R and P aspects as separate components the orders were also inconsistent.

Tentative orders according to the individual R and P components

1st experiment
Writing task:
Receptive Grammar (90%) > Receptive Form and Meaning (88%) = Receptive Associations (88%) > Receptive Orthography (87%) > Productive Grammar (84%) > Receptive Syntax (83%) > Productive Orthography (81%) > Productive Form and Meaning (76%) > Productive Associations (73%) > Productive Syntax (59%)

Reading task:
Receptive Grammar (95%) = Receptive Orthography (95%) > Productive Form and Meaning (94%) > Productive Orthography (92%) > Receptive Form and Meaning (91%) = Receptive Associations (91%) = Productive Grammar (91%) > Productive Syntax (87%) = Productive Associations (87%) > Productive Syntax (73%)

2nd experiment, order for individual R and P components
Writing task:
Receptive Grammar (67%) > Receptive Orthography (65%) > Receptive Associations (61%) > Receptive Syntax (58%) > Receptive Form and Meaning (45%) > Productive Grammar (41%) > Productive Form and Meaning (37%) > Productive Orthography (36%)> Productive Associations (34%)> Productive Syntax (25%).

Reading task:
Receptive Grammar (59%) > Receptive Orthography (53%) > Receptive Syntax (50%) > Receptive Associations(49%) > Productive Grammar (28%) > Receptive Form and Meaning (27%) > Productive Orthography (26%) > Productive Form and Meaning (21%) > Productive Associations (20%) > Productive Syntax (15%).

Results vary depending on the treatment (writing and reading) as well as the time to complete the tasks. When R and P individual aspects are examined, the orders are even less consistent. Therefore, even with the same measures, results seem inconclusive.

Webb (2007a)

Author's conclusion on the degree of knowledge:
Gains in knowledge were found after each encounter for at least one aspect. Receptive knowledge is larger than productive knowledge for each aspect.

****

Tentative orders according to mean % scores for each encounter

1 encounter
Receptive order:
Orthography (67%) > Meaning and Form (58%) > Grammar (57%) > Associations (48%) > Syntax (40%)
Productive order:
Orthography (50%) > Grammar (11%) > Syntax (9.1%) > Associations (6.5%) > Meaning and Form (3.5%)

3 encounters
Receptive order:
Orthography (80%) > Grammar (68%) > Meaning and Form (67.7%) > Associations (57%) > Syntax (55%)
Productive order:
Orthographic (60%) > Association (23%) > Grammar (20%) > Syntax (15%) > Meaning and Form (13%)

7 encounters
Receptive order:
Orthography (83%) > Grammar (70%) > Meaning and Form (65%) > Associations (56%) > Syntax (55%)
Productive order:
Orthography (72%) > Meaning and Form (27%) > Associations (24%) > Syntax (23%) > Grammar (22%)

10 encounters
Receptive order:
Orthography (88%) > Grammar (80%) > Meaning and Form (76%) > Associations (69%) > Syntax (66%)
Productive order:
Orthography (77%) > Grammatical Functions (43%) > Association (40%) > Syntax (37%) > Meaning and Form (29%)

The tentative order of all R and P components based on % of knowledge after 10 encounters would be:
Receptive Orthography (88%) > Receptive Grammar (80%) > Productive Orthography (77%) > Receptive Form–meaning (76%) > Receptive Associations (69%) > Receptive Syntax (66%) > Productive Grammar (43%) > Productive Associations (40%) > Productive Syntax (37%) > Productive Form–meaning (29%)

Webb (2007b)

Author's conclusion on the degree of knowledge:
Scores in the different components in the two treatment conditions (word pairs treatment and context) were not significantly different.

****

Tentative orders according to mean % scores

Word pairs treatment
Receptive Orthography (74%) > Receptive Grammar (69%) > Receptive Associations (60%) > Receptive Syntax (59%) > Productive Form–meaning (56%) = Receptive Form–meaning (56%) > Productive Orthography (55%) > Productive Grammar (49%) > Productive Associations (42%) > Productive Syntax > (41%)

Context treatment
Receptive Orthography (74%) = Receptive Grammar (74%) > Receptive Syntax (67%) > Receptive Associations (65%) > Receptive form–meaning (59%) > Productive Form–meaning (52%) = Productive Orthography (52%) > Productive Grammar (49%) > Productive Associations (44%) > Productive Syntax (38%)

In general, receptive aspects known better than productive aspects.
Orthography the first aspect in the receptive section, and form meaning link the last one.
Easiest in productive part seems to be form–meaning link, followed by orthography, grammar, associations and syntax.

Webb (2009)

Author's conclusion on the degree of knowledge:
Both learning conditions (productive learning and receptive learning) lead to different scores in the components.
Productive task learning lead to better learning in all productive measures and receptive knowledge of orthography. Receptive learning lead to larger gains in receptive form–meaning link.
For most of the aspects the differences seem very small.

****

Tentative orders according to mean % scores for the combined R and P scores in each component

Productive learning task:
Orthography (90%) > Grammar (88%) > Form–meaning (83%) = Associations (83%) > Syntax (72%)

Receptive learning task:
Form–meaning (84%) > Orthography (82%) = Grammar (82%) > Associations (79%) > Syntax (66%)

Tentative orders according to mean % scores for the individual R and P aspects

Productive learning task:
Receptive Orthography (92%) > Receptive Grammar (88%) = Productive Orthography (88%) = Productive Meaning and Form (88%) = Productive Grammar (88%) > Receptive Associations (85%) > Productive Associations (81%) > Receptive Syntax (80%) > Receptive Meaning and Form (78%) > Productive Syntax (62%)

	Receptive learning task: *Receptive Meaning and Form (91%) > Receptive Grammar (90%) > Receptive Associations (87%) > Receptive Orthography (86%) > Receptive Syntax (83%) > Productive Orthography (78%) > Productive Meaning and Form (77%) > Productive Grammar (73%) > Productive Associations (72%) > Productive Syntax (48%)*
Pellicer-Sánchez & Schmitt (2010)	Author's conclusion on the degree of knowledge: Learning gains were not consistent across the word knowledge components from the different exposures. Spelling improves after 5 occurrences, but for the rest of components 10 encounter or more seem to be needed for improvement. Either with 1-8 exposures or 10+ exposures, the scores were highest for meaning recognition, followed by spelling recognition, word class recall and meaning recall. However, there were different proportions of gains. **** Tentative order according to mean % scores after 10 exposures *Meaning Recognition (84%) > Spelling Recognition (76%) > Word class Recall (63%) > Meaning Recall (55%)*
Chen & Truscott (2010)	Author's conclusion on the degree of knowledge: No consistency in results across encounters. Receptive form–meaning link was the worst known aspect after every encounter. Knowledge of orthography, part of speech, and associations showed different patterns of development with increasing encounters. Receptive knowledge of orthography and part of speech seem the easiest aspects. Big distinction/jump in scores between productive and receptive components. **** Tentative orders according to mean % scores ***1 encounter*** Receptive Part of Speech (48%) > Receptive Orthography (43%) > Receptive Associations (40%) > Productive Part of Speech (15%) > Productive Orthography (11%) > Productive Associations (6%) > Receptive Meaning and Form (1%)

3 encounters

Receptive Orthography (63%) > Receptive Part of Speech (55%) > Receptive Associations (39%) > Productive Part of Speech (25%) > Productive Orthography (21%) > Productive Associations (12%) > Receptive Meaning and Form (6%)

7 encounters

Receptive Part of Speech (66%) > Receptive Orthography (65%) > Receptive Associations (51%)> Productive Part of Speech (39%) > Productive Associations (31%) > Productive Orthography (27%) > Receptive Meaning and Form (15%)

Note: > indicates '*easier than'; Italics* = Our tentative order based on the mean percentage scores, not an order claimed by authors.

Appendix 2. Test formats and word knowledge aspects measured in multicomponent studies.

Sample of multicomponents studies, with the components measured, test formats used, and name given to the tests. Our interpretation of what the test is measuring follows Schmitt's (2010) relabelling of Laufer & Goldstein's (2004) categories in an attempt to provide some consistency to the varying terminology.

Components **(Nation's framework)**	**Schmitt & Meara (1997)**	**Schmitt (1998)**	**Webb (2005, 2007a, 2007b, 2009)**	**Chui (2006)**	**Tannenbaum et al. (2006)**	**Pigada & Schmitt (2006)**	**Pellicer-Sánchez & Schmitt (2010)**	**Chen & Truscott (2010)**	**Li & Kirby (2015)**	**Kieffer & Lesaux (2012a)**
Spoken form Productive									**Word Reading Efficiency** *(Form recall)* Pronounce/ read aloud the TWs quickly	
Written form Productive		**Spelling** *(Form recall)* Dictation: Hear a word and spell the word	**Orthography productive knowledge** *(Form recall)* Dictation: Hear a word and write it down			**Spelling** *(Form recall)* Dictation: Hear a word and write it down		**Orthography Productive** *(Form recall)* Dictation: hear a word, and write it down		
Written form Receptive			**Orthography receptive knowledge**				**Spelling Recognition**	**Orthography Receptive**		

		(Form recognition) MC test: circle the correctly spelled word		*(Form recognition)* MC test: select the correct form	*(Form recognition)* MC test: select the correctly spelled word		
Form–meaning Receptive		**Form–meaning receptive knowledge** (2005) *(Meaning recall)* Receptive translation: L2 TW form – L1 equivalent form / (2007a; 2007b; 2009) *(Meaning recognition)* MC test: chose the word that represents the meaning of TW from 4 options (1 correct + 3 distractors)	**PPVT** *(Meaning recognition)* MC test: select the picture that best represents the meaning of a spoken TW	**Recognition of meaning** *(Meaning recognition)* MC test: select the description that best represents meaning of TW from individual cards	**Form–meaning Receptive** *(Meaning recall)* Translation: L2 form – L1 translation/ equivalent	**Gates-MacGinitie Vocabulary Subset** *(Meaning recognition)* MC test: select the closest meaning to a TW	**General synonyms** *(Meaning recognition)* MC task: select a synonym of TWs ****** **Academic synonyms** *(Meaning recognition)* MC test: select synonyms of AWL TWs

Form–meaning Productive		**Form–meaning productive knowledge** (2005) *(Form recall)* Translation: given the L1 meaning produce L2 TW form / **Form–meaning receptive Recall** (2007a; 2007b; 2009) *(Meaning recall)* Translation: L2 TW – L1 equivalent	**Semantic knowledge** *(Meaning recall)* Given the L2 TW, explain the meaning(s) in written (includes multiple meanings)	**WISC Vocabulary Subset** *(Meaning recall)* Define a TW orally ****** **Attributes subset of LPT-R** *(Meaning recall)* Describe TWs, providing different info about their attributes (i.e., function, components, colour, size, shape, category…) (also includes multiple meanings).	**Meanings** *(Meaning recall)* Give any info about meaning(s) of L2 TWs (also multiple meanings)	**Recall of meaning** *(Meaning recall)* Give all info known about the meaning of TWs presented on cards	**Word definitions** *(Meaning recall)* Provide as much info about the meaning of a L2 TW as possible, including category, function, description and value.

Word parts / Derivatives Receptive	**Receptive knowledge of suffixes** *(Form recognition)* MC test with multiple responses: choose from a list of possible inflectional and derivational suffixes the ones allowable for each TW		**Grammatical functions receptive knowledge** *(Form recognition)* MC test: select from 3 options the sentence were the TW is used in the correct PoS	**Grammatical knowledge** *(Word class recognition)* Write the word class / PoS of TWs			**PoS receptive knowledge** *(Form recognition)* MC test: from 3 options, select the sentence in which TW is used in right PoS	**Morphological awareness** *(Word part recognition)* Identify the smallest part (i.e., base) in multi-morphemic words. Word parts recognition	**Real word decomposition (Receptive morphological awareness).** *(Word part recognition)* Extract the base of words with a derivational suffix, provided i oral mode.
Word parts / Derivatives Productive	**Productive knowledge of suffixes** *(Form recall)* Write any inflectional and derivational suffixes allowable for each TW	**Grammatical information recall** *(Word class and derivatives recall)* State the word class of the TWs and how to transform them into the other 3 word classes	**Grammatical functions productive knowledge** *(Form recall)* Sentence construction: write a sentence in which the TW is used in correct PoS	**Morphological Knowledge** *(Form recall)* Provide a derivative form of TWs	**WUF subtest of DIBELS** *(Form recall)* Produce TW in as many sentences per min as possible	**Word class Recall** *(Word class recall)* Write the correct PoS for each TW	**PoS productive knowledge** *(Form recall)* Sentence construction: write a sentence using the TW in the correct PoS		**Nonword derivation (Productive morphological awareness)** *(Form recognition)* Select the nonword from 4 options which ha the appropriate derivational suffi to complete a sentence (e.g., th

						man is a great *tranter*).
Collocation Receptive		**Syntax receptive knowledge** *(Form recognition)* MC test: select, from 4 options, the word most likely to appear in context with the TW	**Collocational knowledge** *(Form recognition)* MC test: choose 1 collocate from 4 options			
Collocation Productive				**Grammar Recall** *(Form recall)* Report any preposition that followed the verb (grammatical collocations)		
Multiple-Meanings Receptive					**Multiple-meaning vocabulary** *(Meaning recognition)* MC test: select the sentence, from 4 options, in	**Multiple meanings** *(Meaning recognition)* MC task: select from 4 sentences the one in which the polysemous TW is used in the

					which the TW is used correctly	same meaning as in the target sentence provided.
Multiple-Meanings Productive		**Meanings** *(Meaning(s) recall)* Interview: Give any info about meaning(s) of TWs	**Multiple-Meanings subset of the LPT-R** *(Meaning recall)* Describe each TW in 3 different contexts, providing synonyms or definition.			
Associations Receptive	**Receptive knowledge of associations** *(Association recognition)* MC with multiple responses: circle the possible associates of a TW from 4 options (3 correct + 1 distractor)	**Association receptive knowledge** *(Meaning recognition)* MC test: circle the correct paradigmatic association of a TW ***** **Syntax receptive knowledge**		**Association Receptive** *(Meaning recognition)* MC test: choose the correct paradigmatic associate of the TW from 5 options		**Semantic Associations (syntagmatic an paradigmatic)** *(Association recognition)* MC task: choose 3 words from 6 options that represent associations of a TW

			(syntagmatic associations) *(Form recognition)* MC test: select the word most likely to appear in context with the TW (collocation form recognition)		
Associations Productive	**Productive knowledge of associations** *(Association recall)* Write 3 words that are associates of the TW	**Association Recall** *(Association recall)* Interview: provide 3 related words of a TW	**Association productive knowledge** *(Association recall)* Write a paradigmatic associate of the TW **** **Syntax Productive knowledge** *(Association recall)* Provide a syntagmatic	**Semantic category** *(Association recall)* Name as many items as possible within each category (e.g., fruits, animals) (Paradigmatic associations)	**Association Productive** *(Association recall)* Write one paradigmatic associate of the TW.

		associate of the TW	
Grammatical information recognition			**Grammar Recognition** *(Article recognition)* Choose which article (feminine or masculine) goes with the noun TWs

Note: TW = Target Word; MC = Multiple-choice; PoS = Part of Speech; PPVT-III = Peabody Picture Vocabulary Test-Third Edition (Dunn & Dunn, 1997); WUF subset of DIBELS = Word Use Fluency subtest of the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002); WISC-III = Wechsler Intelligence Scale for Children–Third Edition (Wechsler, 1991); LPT-R = Language Processing Test-Revised (Richard & Hanner, 1995).

Bold = Author's test label; *Italics* = our interpretation of what knowledge level is being tested; Regular = description and instructions for the actual test.

References

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–Third Edition*. Circle Pines, MN: AGS Publishing.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills*. Eugene, OR: Institute for the Development of Educational Achievement.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition*. New York: Psychological Corporation.

Richard, G. J., & Hanner, M. A. (1995). *Language Processing Test–Revised*. East Moline, IL: LinguiSystems, Inc.

MacGinitie, W. H. & MacGinitie, R. K. (1992). *Gates–MacGinitie Reading Tests*, 2nd Canadian ed. Nelson.

Appendix 3. Complete test battery with background questionnaire.

1. Remember the form

In this first section, you will find a context written in Spanish and then a sentence in English which summarizes or completes that context. Write the appropriate word in the space using the information given or implicit in the context, so that the sentence in Spanish means the same as the sentence in English. You will only need **one** word for each question (one word in each blank). The first letter of each word is shown to help you.

Example:

En la excursión a Paris subimos a lo alto de la Torre Eiffel, pero yo no pude mirar hacia abajo, porque me daba muchísimo miedo. Because I was very scared, the tourist guide told me not to look d___own____.

1.	*Hay que enseñar a los niños y niñas que la luz roja para los peatones en los semáforos quiere decir que hay que parar.* For their safety, we need to teach the children what it m_____.
2.	*La semana pasada me compré unas gafas nuevas, pero los cristales no eran lo suficientemente duros y se me rompieron enseguida.* They weren´t very h_____.
3.	*Cristina lleva ya muchos años contratada como profesora en un colegio.* She has been e_____ for many years now.
4.	*Mi vecino me invitó a su casa a tomar un café. Aunque tenía café de antes, hizo café recién hecho para mí.* My neighbour made a f_____ one for me.
5.	*María siempre ha sido una mujer con mucho encanto y personalidad. Todo el que la conoce se siente inmediatamente atraído por ella.* María has a lot of ch_____ [**NOT** 'charisma'].

6.	*Para poder publicar el artículo en esa revista, es necesario poner sangría en cada nuevo párrafo.* In order to publish, you need to i_____ each new one.
7.	*Por fin, después de muchos años, mi madre va a cumplir su sueño de dar la vuelta al mundo el año que viene.* She is going to f_____ that dream next year.
8.	*Dicen que el niño ha heredado el modo de ser y temperamento de su madre.* He has the ch_____ of his mum.
9.	*El partido de fútbol había sido muy malo, pero el delantero lo compensó en la segunda parte con dos goles fantásticos.* He r_____ himself in the second half of the game.
10.	*Existen distintas fases en el desarrollo de los niños, y hay que conocerlas para adaptarse a ellas.* There are different stages in their d_____.
11.	*El día de mi cumpleaños, mi pareja me pidió que cerrase los ojos, y me sorprendió con un regalo maravilloso.* He asked me to c_____ them to give me a surprise.
12.	*El verano es el mejor periodo del año para mí, porque me gusta mucho el calor y poder ir a la playa.* It is my favorite s_____.
13.	*Ahora mismo no llevo dinero encima para pagarte el dinero de la entrada para el partido, pero voy a sacar dinero a mediodía y te lo devuelvo.* I am going to the b_____ at lunch time.

14.	*El paciente herido en el accidente de tráfico está gravemente enfermo. Los médicos han dicho que no hay cura posible y que no tardará en morirse.* He is in t_____ condition.
15.	*Al entrar en la habitación, tuve que taparme los ojos porque la luz era muy brillante.* It was too b_____ for my eyes. [**NOT** 'brilliant'].
16.	*Hay muchas diferencias entre el inglés de América y el inglés de Reino Unido, sobre todo en el vocabulario.* There is a clear d_____ between them. [**NOT** 'difference' or 'divergence'].
17.	*El profesor me mandó cerrar la ventana porque entraba corriente en la clase.* There was a d_____ in the classroom.
18.	*El año que viene me he propuesto escalar el Everest. Ya sé que es todo un desafío, pero me estoy preparando mucho.* It is a big ch_____.
19.	*Las calles de la ciudad de Nueva York son muy anchas.* The streets are very b_____.
20.	*A Juan no le gusta la zanahoria, así que su padre se la ralla para que se la coma.* Juan's dad always g_____ the carrot for him.

2. Remember the meaning

Choose the correct meaning of the underlined words in *italics* and circle how sure you are about your answer (**1** = not sure at all; **2** = not very sure; **3** = almost sure; **4** = totally sure). See Example 1.

Example 1:

		How sure are you of your answer?			
1.	He was told to get *down*. a) Get to a higher place b) Get to a lower place c) Go to the first place d) Go to the last place e) I don't know	1	2	3	4

If you don't know the answer at all, select the "I don't know" option, and don't say how secure you are of the answer. See Example 2.

Example 2:

		How sure are you of your answer?			
2.	I left the plates in the *sink*. a) Place to show things b) Place to store things c) Place to fix things d) Place to wash things e) I don't know	1	2	3	4

How sure are you of your answer?
1 = not sure at all **2** = not very sure **3** = almost sure **4** = totally sure

		How sure are you of your answer?			
1.	I know what it *means*. a) Looks like b) Is made up of c) Indicates d) Becomes later e) I don't know	1	2	3	4
2.	I don't like your *character*. a) Attitude b) Clothes c) Pet d) Opinion e) I don't know	1	2	3	4
3.	You need to *indent* the text. a) Read carefully b) Improve the spelling c) Add to the ideas d) Move to the right e) I don't know	1	2	3	4
4.	The house had some *charm*. a) Value b) Furniture c) Good location d) Nice feeling e) I don´t know	1	2	3	4
5.	He wants to *fulfil* his hope. a) Start b) Forget c) Complete d) Change e) I don´t know	1	2	3	4

		How sure are you of your answer?			
6.	She is *employed*. a) She is lazy b) She is busy c) She has a job d) She has a lot of time e) I don't know	1	2	3	4
7.	We need some *fresh* evidence. a) Familiar b) Original c) Careful d) Popular e) I don't know	1	2	3	4
8.	He wanted to *redeem* himself. a) Compensate himself for a good action b) Cure his injury c) Make up for a mistake d) Improve his knowledge e) I don't know	1	2	3	4
9.	The *development* is good. a) Game b) Progress c) Animal d) Feeling e) I don't know	1	2	3	4
10.	The material is *hard*. a) Cheap b) Dirty c) Broken d) Solid e) I don't know	1	2	3	4

		How sure are you of your answer?			
1.	His situation is *terminal.* a) He is healthy b) He is sick c) He will soon die d) Doctors don't know e) I don't know	1	2	3	4
2.	It is too *bright* outside. a) Sunny b) Dark c) Hot d) Cold e) I don´t know	1	2	3	4
3.	I only like to *grate* vegetables. a) Cook slowly b) Cut into small pieces c) Grill in the oven d) Boil in water e) I don't know	1	2	3	4
4.	I don't like that *bank.* a) Place to sit b) Short dress c) Place for money d) Food from Greece e) I don't know	1	2	3	4
5.	It was very *broad*. a) Wide b) Long c) Deep d) Heavy e) I don't know	1	2	3	4

		How sure are you of your answer?			
6.	It is the best *season.* a) Animal b) Time c) Appearance d) Place e) I don't know	1	2	3	4
7.	That is a *challenge*. a) Something difficult b) Something made at home c) Something magical d) Something from the sea e) I don't know	1	2	3	4
8.	There is a *draught* in here. a) Small insect b) Soft light c) Small wind d) Cold temperature e) I don't know	1	2	3	4
9.	I want you to *close* it. a) Paint b) Clean c) Shut d) Buy e) I don't know	1	2	3	4
10.	There was a *distinction* between them. a) Argument b) Problem c) Difference d) Person e) I don't know	1	2	3	4

3. Word class formation

Most words in English can be changed into different words with different word classes (this is, noun, verb, adjective or adverb). For example, STIMULATE is a verb in English, but it can be changed into a noun → STIMULATION; or an adjective → STIMULATING. In this case, STIMULATE does not have an adverb form, but many other words have all the forms. In other cases, the form can be the same in different word classes. For example, BENEFIT is a noun, and BENEFIT is also the verb.

In this section, write the form of the word in *italics* that corresponds to the noun, verb, adjective and adverb in each sentence. You only need **one** word for each gap. If there is more than one possibility (e.g., more than one adjective form) you only need to write the one that fits the best in the context. If there is no form for a word class, put an "X" in the blank. Sometimes the word in *italics* will not need changing, as it might already be correct (such as STIMULATE in the following example). Please, **do not guess**. If you do not know the answer, just leave it blank.

Example 1:

Stimulate

Noun	A massage gives good ___stimulation_____.
Verb	Massages can _____stimulate_____ tired muscles.
Adjective	A massage has a _____stimulating___ effect.
Adverb	He massaged ___X___.

Example 2:

Benefit

Noun	The __benefit__ of charity centres is happier children.
Verb	Charity centres organize activities that __benefit__ children.
Adjective	Charity centres are __beneficial__.
Adverb	Charity centres organize children´s activities __beneficially__.

1.	***Development***	
	Noun	The _____ of a person depends on everything that is around them.
	Verb	The environment is very important for a person to _____ well.
	Adjective	The environment is very important in the _____ stages of a person.
	Adverb	A good environment is _____ important for children.
2.	***Charm***	
	Noun	The _____ of the old cathedral impressed me a lot.
	Verb	The old cathedral could _____ anyone.
	Adjective	The old cathedral was very _____.
	Adverb	The cathedral was _____ designed.
3.	***Character***	
	Noun	Politeness is traditionally part of the British _____.
	Verb	Politeness is said to _____ British people.
	Adjective	Politeness is traditionally _____ of British people.
	Adverb	Politeness _____ describes British people.
4.	***Employ***	
	Noun	I am very happy to gain _____ in this company.
	Verb	This company will _____ 150 people when it grows.
	Adjective	My good qualifications made me very _____ at this company.
	Adverb	I want to work _____ at this company.

5.	***Redeem***	
	Noun	The _____ of the winning lottery ticket will pay £1 million.
	Verb	He has to _____ the winning lottery ticket.
	Adjective	The winning lottery ticket is _____ in any shop.
	Adverb	He got the winning lottery ticket _____
6.	***Hard***	
	Noun	The _____ of steel makes it a good building material.
	Verb	The company must _____ its steel to make it stronger.
	Adjective	Steel is a _____ material.
	Adverb	The steel company _____ produces waste.
7.	***Mean***	
	Noun	The _____ of these reports is clear; our company is in big trouble.
	Verb	These reports _____ that some people will be fired.
	Adjective	These reports are _____, because they will change the company´s future.
	Adverb	We can _____ compare these reports with last year´s reports, and see that our company is getting worse.
8.	***Fulfil***	
	Noun	My job gives me a sense of _____.
	Verb	My job helps me to _____ my hopes.
	Adjective	My job is very _____.
	Adverb	I do my job _____.

9.	***Indent***	
	Noun	You need to make an _____ for each quote.
	Verb	You need to _____ each quote.
	Adjective	You need to edit the _____ quotes.
	Adverb	You need to edit the quote _____.
10.	***Fresh***	
	Noun	I enjoy the _____ of the early morning in summer days.
	Verb	The wind will always _____ in the early morning in summer days.
	Adjective	The wind is very _____ in the early morning in summer days.
	Adverb	The wind blows _____ in the early morning in summer days.
11.	***Terminal***	
	Noun	Because of the baby´s disorder, she agreed to the _____ of her pregnancy.
	Verb	Because of the baby´s disorder, she decided to _____ her pregnancy.
	Adjective	The baby had a _____ disorder.
	Adverb	The baby was _____ ill.
12.	***Challenge***	
	Noun	Finding a cure for cancer is the biggest _____ in medicine today.
	Verb	Finding a cure for cancer will _____ medical researchers for many years.
	Adjective	Finding a cure for cancer is a _____ job.
	Adverb	Medical researchers are working _____ to find a cure for cancer.

13.	***Grate***	
	Noun	Our _____ broke, so I had to buy a new one for the carrots.
	Verb	Carrots taste better when you _____ them on top of a salad.
	Adjective	_____ carrots are good on top of a salad.
	Adverb	I always serve the carrots _____ for salads.
14.	***Bright***	
	Noun	My mum will like the _____ of my living room.
	Verb	The large number of lights will _____ the living room.
	Adjective	My mum will like my _____ living room.
	Adverb	My living room will be lit _____.
15.	***Distinction***	
	Noun	Marcos and Pedro are twins; there is almost no _____ between them.
	Verb	It is very difficult to _____ between Marcos and Pedro, because they are very similar.
	Adjective	There is no _____ feature between Marcos and Pedro.
	Adverb	Nobody can _____ say who is Marcos and who is Pedro.
16.	***Season***	
	Noun	In this country, each _____ is clearly different.
	Verb	In this country, the temperature variations _____ clearly.
	Adjective	In this country, the _____ temperature variations are clearly different.
	Adverb	In this country, the temperature variations occur _____.

17.	***Close***	
	Noun	The _____ of my children's school will happen in July.
	Verb	My children's school will _____ for ever in July.
	Adjective	My children's school will be _____ from August on.
	Adverb	My children's school will stop working _____.
18.	***Broad***	
	Noun	The _____ of the road was impressive.
	Verb	We didn't need to _____ the road; it was already wide enough.
	Adjective	When we arrived, we found that the road was very _____.
	Adverb	The road extended _____ in front of us.
19.	***Draught***	
	Noun	The window was open and there was a _____ in the room.
	Verb	I opened the window to_____ the room.
	Adjective	The room was _____ because the window was open.
	Adverb	I opened the window _____.
20.	***Bank***	
	Noun	My family always goes to the same _____.
	Verb	My family likes to _____ with just one company.
	Adjective	My mum is a very _____ Hollywood star.
	Adverb	My mum always goes to that company because the interest is _____ high.

4. Word class recognition

In this part, given the word in **bold**, choose the appropriate word (form) that corresponds to the noun, verb, adjective and adverb in each sentence. Choose the letter from the options (***a-h***) that represents that word and write it in the blank.

There are some incorrect forms among the options given, so be careful. In **Example 1**, *stimulize*, *stimulatement*, *stimulaty* and *stimulatable* are words that do not exist in English. Some words may not have a specific form for one or more of their word classes (for example, not having a form for the verb). If there is no form (X), choose the option "***h***". Then, select how sure you are of your answers from 1 to 4 (**1** = not sure at all; **2** = not very sure; **3** = almost sure; **4** = totally sure). Please, **do not guess**. If you do not know the answer, just leave it blank.

Example 1:

Stimulate

a. *Stimulize*	***b.*** *Stimulate*	***c.*** *Stimulaty*	***d.*** *Stimulating*
e. *Stimulatement*	***f.*** *Stimulation*	***g.*** *Stimulatable*	***h.*** *X*

		How sure are you of your answers?
Noun	A massage gives good ___**f**_____.	1 2 3 (4)
Verb	Massages can _____**b**_____ tired muscles.	1 2 (3) 4
Adjective	A massage has a _____**d**___ effect.	1 2 3 (4)
Adverb	He massaged ___**h**___.	(1) 2 3 4

In some cases, you may need to use the same form two or more times. In the following example, the form *"benefit"* can be used both as a noun and as a verb.

Example 2:

a. *Benefit*	***b.*** *Beneficially*	***c.*** *Benefate*	***d.*** *Benefist*
e. *Beneficial*	***f.*** *Benefary*	***g.*** *Benefarily*	***h.*** *X*

Benefit

		How sure are you of your answers?
Noun	Child activities have the __**a**__ of support from charity centers.	1 2 3 4
Verb	Charity centers support activities that __**a**__ children.	1 2 3 4
Adjective	Charity centers are __**e**__.	1 2 3 4
Adverb	Charity centers support children activities __**b**__.	1 2 3 4

1.

Fresh

a. *Freshen*	***b.*** *Fresh*	***c.*** *Freshable*	***d.*** *Freshness*
e. *Freshment*	***f.*** *Freshate*	***g.*** *Freshly*	***h.*** *X*

		How sure are you of your answers?			
Noun	I enjoy the _____ of the early morning in summer days.	1	2	3	4
Verb	The wind will always _____ in the early morning in summer days.	1	2	3	4
Adjective	The wind is very _____ in the early morning in summer days.	1	2	3	4
Adverb	The wind blows _____ in the early morning in summer days.	1	2	3	4

2.

Indent

a. *Indentment*	***b.*** *Indentive*	***c.*** *Indentify*	***d.*** *Indentedly*
e. *Indented*	***f.*** *Indent*	***g.*** *Indentively*	***h.*** *X*

		How sure are you of your answers?			
Noun	You need to make an _____ for each quote.	1	2	3	4
Verb	You need to _____ each quote.	1	2	3	4
Adjective	You need to edit the _____ quotes.	1	2	3	4
Adverb	You need to edit the quote _____.	1	2	3	4

3.

Character

a. *Characterous*	***b.*** *Characterness*	***c.*** *Character*	***d.*** *Characterify*
e. *Characteristically*	***f.*** *Characteristic*	***g.*** *Characterize*	***h.*** *X*

		How sure are you of your answers?			
Noun	Politeness is traditionally part of the British _____.	1	2	3	4
Verb	Politeness is said to _____ British people.	1	2	3	4
Adjective	Politeness is traditionally _____ of British people.	1	2	3	4
Adverb	Politeness _____ describes British people.	1	2	3	4

4.

Hard

a. *Hard*	***b.*** *Hardful*	***c.*** *Hardness*	***d.*** *Hardify*
e. *Harden*	***f.*** *Hardly*	***g.*** *Hardity*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of steel makes it a good building material.	1	2	3	4
Verb	The company must _____ its steel to make it stronger.	1	2	3	4
Adjective	Steel is a _____ material.	1	2	3	4
Adverb	The steel company _____ produces waste.	1	2	3	4

5.

Redeem

a. *Redeem*	***b.*** *Redeemit*	***c.*** *Redeemable*	***d.*** *Redeemably*
e. *Redemption*	***f.*** *Redeemity*	***g.*** *Redeemal*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of the winning lottery ticket will pay £1 million.	1	2	3	4
Verb	He has to _____ the winning lottery ticket.	1	2	3	4
Adjective	The winning lottery ticket is _____ in any shop.	1	2	3	4
Adverb	He got the winning lottery ticket _____.	1	2	3	4

6.

Development

a. *Develop*	***b.*** *Developate*	***c.*** *Developmental*	***d.*** *Developable*
e. *Developmentally*	***f.*** *Development*	***g.*** *Developness*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of a person depends on everything that is around them.	1	2	3	4
Verb	The environment is very important for a person to _____ well.	1	2	3	4
Adjective	The environment is very important in the _____ stages of a person.	1	2	3	4
Adverb	A good environment is _____ important for children.	1	2	3	4

7.

Employ

a. *Employ*	***b.*** *Employance*	***c.*** *Employable*	***d.*** *Employably*
e. *Employment*	***f.*** *Employify*	***g.*** *Employal*	***h.*** *X*

		How sure are you of your answers?			
Noun	I am very happy to gain _____ in this company.	1	2	3	4
Verb	This company will _____ 150 people when it grows.	1	2	3	4
Adjective	My good qualifications made me very _____ at this company.	1	2	3	4
Adverb	I want to work _____ at this company.	1	2	3	4

8.

Charm

a. *Charmness*	***b.*** *Charmate*	***c.*** *Charming*	***d.*** *Charm*
e. *Charmfully*	***f.*** *Charmingly*	***g.*** *Charmful*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of the old cathedral impressed me a lot.	1	2	3	4
Verb	The old cathedral could _____ anyone.	1	2	3	4
Adjective	The old cathedral was very _____.	1	2	3	4
Adverb	The cathedral was _____ designed.	1	2	3	4

9.

Fulfil

a. *Fulfilingly*	***b.*** *Fulfilness*	***c.*** *Fulfilize*	***d.*** *Fulfiling*
e. *Fulfil*	***f.*** *Fulfilment*	***g.*** *Fulfilish*	***h.*** *X*

		How sure are you of your answers?			
Noun	My job gives me a sense of _____.	1	2	3	4
Verb	My job helps me to _____ my hopes.	1	2	3	4
Adjective	My job is very _____.	1	2	3	4
Adverb	I do my job _____.	1	2	3	4

10.

Mean

a. *Meaningable*	***b.*** *Meaningful*	***c.*** *Meaningness*	***d.*** *Meaning*
e. *Mean*	***f.*** *Meanify*	***g.*** *Meaningfully*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of these reports is clear; our company is in big trouble.	1	2	3	4
Verb	These reports _____ that some people will be fired.	1	2	3	4
Adjective	These reports are _____, because they will change the company´s future.	1	2	3	4
Adverb	We can _____ compare these reports with last year´s reports, and see that our company is getting worse.	1	2	3	4

11.

Bright

a. *Brightly*	***b.*** *Bright*	***c.*** *Brightify*	***d.*** *Brighten*
e. *Brightness*	***f.*** *Brightment*	***g.*** *Brightful*	***h.*** *X*

		How sure are you of your answers?			
Noun	My mum will like the _____ of my living room.	1	2	3	4
Verb	The large number of lights will _____ the living room.	1	2	3	4
Adjective	My mum will like my _____ living room.	1	2	3	4
Adverb	My living room will be lit _____.	1	2	3	4

12.

Close

a. *Closen*	***b.*** *Closure*	***c.*** *Close*	***d.*** *Closed*
e. *Closement*	***f.*** *Closive*	***g.*** *Closely*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of my children's school will happen in July.	1	2	3	4
Verb	My children's school will _____ for ever in July.	1	2	3	4
Adjective	My children's school will be _____ from August on.	1	2	3	4
Adverb	My children's school will stop working _____.	1	2	3	4

13.

Draught

a. *Draughty*	***b.*** *Draughtness*	***c.*** *Draughted*	***d.*** *Draught*
e. *Draughtedly*	***f.*** *Draughtify*	***g.*** *Draughtily*	***h.*** *X*

		How sure are you of your answers?			
Noun	The window was open and there was a _____ in the room.	1	2	3	4
Verb	I opened the window to _____ the room.	1	2	3	4
Adjective	The room was _____ because the window was open.	1	2	3	4
Adverb	I opened the window _____.	1	2	3	4

14.

Broad

a. *Breadth*	***b.*** *Broadment*	***c.*** *Broaden*	***d.*** *Broadful*
e. *Broadly*	***f.*** *Broadify*	***g.*** *Broad*	***h.*** *X*

		How sure are you of your answers?			
Noun	The _____ of the road was impressive.	1	2	3	4
Verb	We didn't need to _____ the road; it was already wide enough.	1	2	3	4
Adjective	When we arrived, we found that the road was very _____.	1	2	3	4
Adverb	The road extended _____ in front of us.	1	2	3	4

15.

Grate

a. *Grate*	***b.*** *Grated*	***c.*** *Grater*	***d.*** *Gratedly*
e. *Grateful*	***f.*** *Grateness*	***g.*** *Gratate*	***h.*** *X*

		How sure are you of your answers?			
Noun	Our _____ broke, so I had to buy a new one for the carrots.	1	2	3	4
Verb	Carrots taste better when you _____ them on top of a salad.	1	2	3	4
Adjective	_____ carrots are good on top of a salad.	1	2	3	4
Adverb	I always serve the carrots _____ for salads.	1	2	3	4

16.

Season

a. *Season*	***b.*** *Seasonize*	***c.*** *Seasonally*	***d.*** *Seasonation*
e. *Seasonate*	***f.*** *Seasonal*	***g.*** *Seasony*	***h.*** *X*

		How sure are you of your answers?			
Noun	In this country, each _____ is clearly different.	1	2	3	4
Verb	In this country, the temperature variations _____ clearly.	1	2	3	4
Adjective	In this country, the _____ temperature variations are clearly different.	1	2	3	4
Adverb	In this country, the temperature variations occur _____.	1	2	3	4

17.

Challenge

a. *Challengize*	***b.*** *Challenge*	***c.*** *Challenging*	***d.*** *Challengefully*
e. *Challengingly*	***f.*** *Challengeful*	***g.*** *Challengeship*	***h.*** *X*

		How sure are you of your answers?			
Noun	Finding a cure for cancer is the biggest _____ in medicine today.	1	2	3	4
Verb	Finding a cure for cancer will _____ medical researchers for many years.	1	2	3	4
Adjective	Finding a cure for cancer is a _____ job.	1	2	3	4
Adverb	Medical researchers are working _____ to find a cure for cancer.	1	2	3	4

18.

Bank

a. *Bankness*	***b.*** *Bankate*	***c.*** *Bankably*	***d.*** *Bank*
e. *Bankable*	***f.*** *Bankedly*	***g.*** *Bankful*	***h.*** *X*

		How sure are you of your answers?			
Noun	My family always goes to the same _____.	1	2	3	4
Verb	My family likes to _____ with just one company.	1	2	3	4
Adjective	My mum is a very _____ Hollywood star.	1	2	3	4
Adverb	My mum always goes to that company because the interest is _____ high.	1	2	3	4

19.

Terminal

a. *Terminal*	***b.*** *Terminance*	***c.*** *Terminous*	***d.*** *Terminize*
e. *Terminate*	***f.*** *Terminally*	***g.*** *Termination*	***h.*** *X*

		How sure are you of your answers?			
Noun	Because of the baby´s disorder, she agreed to the _____ of her pregnancy.	1	2	3	4
Verb	Because of the baby´s disorder, she decided to _____ her pregnancy.	1	2	3	4
Adjective	The baby had a _____ disorder.	1	2	3	4
Adverb	The baby was _____ ill.	1	2	3	4

20.

Distinction

a. *Distinctly*	***b.*** *Distinctable*	***c.*** *Distinctify*	***d.*** *Distinguish*
e. *Distinctment*	***f.*** *Distinction*	***g.*** *Distinct*	***h.*** *X*

		How sure are you of your answers?			
Noun	Marcos and Pedro are twins; there is almost no _____ between them.	1	2	3	4
Verb	It is very difficult to _____ between Marcos and Pedro, because they are very similar.	1	2	3	4
Adjective	There is no _____ feature between Marcos and Pedro.	1	2	3	4
Adverb	Nobody can _____ say who is Marcos and who is Pedro.	1	2	3	4

5. Polysemy recall

A lot of words in English are polysemous, that means, they have several meanings. All the words in this section are polysemous, and thus have different meanings. Some of these meanings are clearer than others.

To answer this part of the test you need to provide information about all the different meanings you know of the words in *italics*. You can give a definition, or a translation, or synonyms, or opposites, or use them in a correct sentence in that context, or any information that clearly shows that you know the different meanings. In order to help you, there are some prompts for each meaning written in brackets. But be careful, if you **only** provide a translation of the clues given, the answer will be considered incorrect.

Example:

Down

(Adverb= position/level): abajo/ they walked down the mountain

(Adjective= feeling) estar de bajón/feeling unhappy, without energy

(Noun= birds/hair) plumón or pelusa/soft feathers of a bird, or soft hair

1.	***Charm*** (Noun= kind of quality)_____ (Noun= object)_____ (Verb/Noun= fairy tale)_____
2.	***Hard*** (Adjective= hard table)_____ (Adjective= hard exam)_____ (Adjective= with effort)_____
3.	***Fulfil*** (Verb= fulfil a wish)_____ (Verb= obligation)_____ (Verb= happy)_____
4.	***Fresh*** (Adjective= ideas)_____ (Adjective= weather)_____ (Adjective= water)_____

5.	***Character*** (Noun= describing someone)_____ (Noun= in a story)_____ (Noun= character in writing)_____
6.	***Redeem*** (Verb= done to past bad behaviour)_____ (Verb= done to a coupon)_____ (Verb= when you are in debt)_____
7.	***Development*** (Noun= of a person)_____ (Noun= of new things)_____ (Noun= houses)_____
8.	***Employ*** (Verb=to pay for something)_____ (Verb= for a purpose)_____ (Verb= employ time)_____

9.	***Mean*** (Verb= describing something)_____ (Noun= maths)_____ (Adjective= attitude)_____
10.	***Indent*** (Verb= when writing)_____ (Verb= indent for equipment)_____ (Noun= in a car accident)_____
11.	***Terminal*** (Adjective= illness)_____ (Noun= train)_____ (Noun= electronic machine)_____
12.	***Grate*** (Verb= cooking)_____ (Verb= teeth)_____ (Verb= how loud people make you feel)_____

13.	***Challenge*** (Noun= describing a situation)_____ (Verb= done in a competition)_____ (Verb= to challenge the truth)_____
14.	***Distinction*** (Noun= distinction between two things)_____ (Noun= at university)_____ (Noun= something special)_____
15.	***Season*** (Noun= year)_____ (Verb= cooking)_____ (Noun= animals in season)_____
16.	***Close*** (Verb= door)_____ (Verb= a meeting)_____ (Adjective= a close place)_____

17.	***Bank*** (Noun= money)_____ (Noun= water)_____ (Verb=plane)_____
18.	***Draught*** (Noun= wind)_____ (Noun= way of serving drinks)_____ (Noun= game)_____
19.	***Bright*** (Adjective= light)_____ (Adjective= learning)_____ (Adjective= future of something)_____
20.	***Broad*** (Adjective= describing a river)_____ (Adjective= variety)_____ (Adjective= way of speaking)_____

6. Collocation recall

Some words occur together in a language more common than others. These are called collocations. In this section, you will find a context in Spanish and then a sentence in English. Write a word in the gap using the information given or implicit in the context, so that the word makes sense in the context and sounds natural together with the word underlined and in *italics*. You will only need **one** word for each question (one word in each blank), and the first letter of each word is shown to help you.

Example 1:

Mi profesor me dijo que me tranquilizara.

The teacher told me to *calm* d_own_____ because I was really stressed.

Example 2:

Meterse en un buen lío.

When Peter´s dad found out he had broken the TV, he told him: "you are in s_erious__ *trouble*".

1.	*Poner sangría al comienzo de un texto.* When writing, you should *indent* the first l_____ of a text.
2.	*Su padre les aseguró que les llevaría al cine, pero no cumplió su palabra.* Their dad did not *fulfill* his p_____ to take them to the movies.
3.	*Oxigenar el ambiente.* It was too warm in the classroom, so I asked the teacher to let some *fresh* a_____ into the classroom.

4.	*Fomentar buenos hábitos para mantener el medio ambiente.* Factories and big companies need to take more care about the environment and promote s_____ *development*.
5.	*Trabajar pocas horas al día.* He is only *employed* p_____ -_____ in the company, just 6 hours per week.
6.	*La valoración media es 9.25* The *mean* s_____ of all her tests in the competition was 9.25.
7.	*Difícil tarea.* Taking care of babies is very *hard* w_____.
8.	*Personajes más importantes.* There are three m_____ *characters* in this film: the mother, the daughter and her brother.
9.	*Sin ellos te sientes desprotegido y temes infortunios.* He didn't get on the train because he had forgotten his l_____ *charm* and felt afraid of something happening.
10.	*Enmendar problemas hasta recuperar el prestigio y la popularidad.* Pedro didn't have many friends because he had stolen some money. Eventually, he managed to make up for his mistake and *redeem* his r_____.
11.	*Gran oferta de titulaciones.* This University offers a *broad* r_____ of degrees, including sciences, arts and law.
12.	*Marta y Andrea siempre han estado muy unidas, son compañeras íntimas.* Marta and Andrea are very *close* f_____. They are always together.

13.	*En los bares se sirve así, y hay gente que opina que sabe mejor.* Every time we go to a pub, my cousin orders *draught* b_____, because he doesn't like the bottled one.
14.	*Pedir dinero en el banco.* Pedro asked for a *bank* l_____ to get some money to expand his own business.
15.	*Enfrentarse a un reto.* Most young people today have to f_____ the great *challenge* of going to university and study a degree.
16.	*Los tonos vivos hacen sentir bien.* I like the paintings with a lot of *bright* c_____ in them. They make me feel happy.
17.	*Establecer diferencias.* In teaching, it is important to m_____ a *distinction* between oral and written language.
18.	*Condiciones de la salud que causan la muerte.* There is a lot of research in medicine to fight against *terminal* d_____.
19.	*En temporada alta es cuando más gente se va de vacaciones.* When you plan to go on holidays you should bear in mind that hotels are always more expensive in p_____ *season*.
20.	*Es más barato comprarlo entero y rallarlo en casa. Necesario en pizzas.* I always *grate* the ch_____ at home to put on the pizzas.

7. Polysemy recognition

The words in *italics* in this section have different meanings. Each word is presented in 5 different situations, with a different meaning in each situation. Choose the sentences which are correct. Some are correct and some are not correct, but there is always between one and three correct sentences.

Then, select how sure you are of your answers from 1 to 4 (1 = not sure at all; 2 = not very sure; 3 = almost sure; 4 = totally sure). Please, **do not guess**. If you do not know any meaning, select the "I don't know" option, and do not circle how sure you are of your answer.

Example:

	How sure are you of your answer?
Down	
a) I bent down to look under the table.	1 2 3 **4**
b) He has been down all week.	1 **2** 3 4
c) The down of my car broke.	1 2 3 4
d) I wear down jackets all winter.	**1** 2 3 4
e) I have a down on my guitar.	1 2 3 4
f) I don't know	

		How sure are you of your answer?			
1.	***Fresh***				
	a) After I had finished the work, my boss gave me fresh instructions.	1	2	3	4
	b) He bought a totally fresh car.	1	2	3	4
	c) The window glass was very fresh that morning.	1	2	3	4
	d) These plants only survive in fresh water.	1	2	3	4
	e) The mornings are always very fresh in the mountains.	1	2	3	4
	f) I don't know.				

2.	***Redeem***				
	a) She decided to volunteer to redeem herself.	1	2	3	4
	b) The money required to redeem the mortgage was very high.	1	2	3	4
	c) The ticket can be redeemed at any store of this shopping centre.	1	2	3	4
	d) I bought myself a coffee to redeem my progress.	1	2	3	4
	e) The doctor told her to rest in order to redeem the pain.	1	2	3	4
	f) I don't know				
3.	***Employ***				
	a) Matt has been employed for over 2 years now.	1	2	3	4
	b) She employs her days in gardening and playing cards.	1	2	3	4
	c) The purpose was to employ to his boss about his low salary.	1	2	3	4
	d) They have employed new methods in their research.	1	2	3	4
	e) Workers employ in the school from 9 to 5 pm, Monday to Friday.	1	2	3	4
	f) I don´t know.				
4.	***Fulfil***				
	a) You need a black pen to fulfil the form.	1	2	3	4
	b) The building does not fulfil any more people.	1	2	3	4
	c) You need to fulfil your goals to stay happy.	1	2	3	4
	d) The policy does not fulfil any need in the community.	1	2	3	4
	e) Her daughter fulfils her.	1	2	3	4
	f) I don´t know				

5.	***Charm***				
	a) Joe´s charm was scared and run away.	1	2	3	4
	b) Many people carry charms for protection against evil.	1	2	3	4
	c) The town´s charm made me want to return soon.	1	2	3	4
	d) The rude charm wasn´t allowed in the room.	1	2	3	4
	e) The story says the bad witch used magic to charm the children.	1	2	3	4
	f) I don't know				
6.	***Indent***				
	a) The doctor had to indent our meeting to the following week.	1	2	3	4
	b) The Government will indent for new mobile phones next month.	1	2	3	4
	c) There was a small indent on the wall where I shot it.	1	2	3	4
	d) It took me very long to indent all the paragraphs.	1	2	3	4
	e) It is my turn to indent for drinks at the pub.	1	2	3	4
	f) I don't know				
7.	***Mean***				
	a) The dog has a mean number of legs.	1	2	3	4
	b) He is very mean to his brother.	1	2	3	4
	c) The red sign means 'Stop'.	1	2	3	4
	d) In mathematics, the mean is the same as the average.	1	2	3	4
	e) The chair means to be in the corner.	1	2	3	4
	f) I don't know				

8.	***Character***				
	a) This is a character in Chinese writing.	1	2	3	4
	b) With the new head teacher, the character of the school has changed completely.	1	2	3	4
	c) Her character was taken to the doctors.	1	2	3	4
	d) The best character of the play died very early.	1	2	3	4
	e) The character tastes delicious when my mother cooks it.	1	2	3	4
	f) I don't know				
9.	***Hard***				
	a) No matter how hard I tried, I couldn't finish the book.	1	2	3	4
	b) Diamonds are very hard.	1	2	3	4
	c) We bought an original hard cat from India.	1	2	3	4
	d) When you are an adult you have to make hard decisions.	1	2	3	4
	e) The colour of the dress is hard for the shoes.	1	2	3	4
	f) I don't know				
10.	***Development***				
	a) Marcos forgot his development in the supermarket.	1	2	3	4
	b) The film is about the development of music in the last 50 years.	1	2	3	4
	c) She is working on the development of a new computer.	1	2	3	4
	d) I sold the land for a new building development.	1	2	3	4
	e) Please, go and check the development of my phone charging.	1	2	3	4
	f) I don't know				

11.	***Bank***				
	a) We put some money into the bank every week.	1	2	3	4
	b) I forgot the bank of the car at the petrol station.	1	2	3	4
	c) These flowers generally grow on the bank of the river.	1	2	3	4
	d) The helicopter started to bank to the left.	1	2	3	4
	e) She always buys a bank of books when she goes to the book shop.	1	2	3	4
	f) I don't know				
12.	***Grate***				
	a) The potatoes need to be grated into a mixing bowl.	1	2	3	4
	b) Grated music is used in hospitals to calm down patients and relatives.	1	2	3	4
	c) Mary always grates her teeth when she is angry.	1	2	3	4
	d) Cook the grate into the microwave for 10 minutes.	1	2	3	4
	e) The way he talks about himself really grates on me.	1	2	3	4
	f) I don´t know				
13.	***Bright***				
	a) The snow was bright on that winter day.	1	2	3	4
	b) He was surprised reading the bright news in the magazine.	1	2	3	4
	c) His daughter is considered to be very bright.	1	2	3	4
	d) Things will be bright for the company next year.	1	2	3	4
	e) The boss felt very proud of his bright situation.	1	2	3	4
	f) I don´t know				

14.	***Terminal***				
	a) The tree was terminal because of the high temperatures.	1	2	3	4
	b) His condition was terminal.	1	2	3	4
	c) This bus will take you to the terminal.	1	2	3	4
	d) Nowadays we cannot live without computer terminals.	1	2	3	4
	e) We always meet at the school´s terminal.	1	2	3	4
	f) I don't know				
15.	***Season***				
	a) The four seasons are winter, spring, summer and autumn.	1	2	3	4
	b) The car´s season breaks very often.	1	2	3	4
	c) Their dog is in season and can´t go out.	1	2	3	4
	d) It is important to check the season of your computer once a year.	1	2	3	4
	e) I forgot to season the fish with salt and pepper.	1	2	3	4
	f) I don´t know				
16.	***Broad***				
	a) I went to the supermarket and bought broad fruit.	1	2	3	4
	b) The bridge was broad and had plenty of space.	1	2	3	4
	c) The show was created for a broad public.	1	2	3	4
	d) The light in the city was very broad.	1	2	3	4
	e) His family have a broad accent.	1	2	3	4
	f) I don't know				

17.	***Draught***				
	a) I prefer to drink it on draught, not from a bottle.	1	2	3	4
	b) I am very good at playing draughts.	1	2	3	4
	c) She took a draught of my sandwich.	1	2	3	4
	d) There was a draught in the director's computer.	1	2	3	4
	e) This plant does not like cold draughts in the room.	1	2	3	4
	f) I don't know				
18.	***Challenge***				
	a) Starting a new company is a big challenge for me.	1	2	3	4
	b) Most people now challenge the decisions of the politicians.	1	2	3	4
	c) The fisherman caught this challenge in the morning.	1	2	3	4
	d) Gary will challenge Matt to a game of tennis.	1	2	3	4
	e) She wants me to challenge her a motorbike for her birthday.	1	2	3	4
	f) I don't know				
19.	***Distinction***				
	a) I bought my bag's distinction in a street market in the USA.	1	2	3	4
	b) My dad had the distinction to celebrate my birthday at the beach.	1	2	3	4
	c) I can't see the distinction between the two phones.	1	2	3	4
	d) She got a distinction for her paper.	1	2	3	4
	e) He has the distinction of being the tallest person in the world.	1	2	3	4
	f) I don't know				

20.	***Close***				
	a) I have to return the book today because the library will close tomorrow.	1	2	3	4
	b) The politician closed the discussion saying thank you to the presenter.	1	2	3	4
	c) On our last holiday, we stayed at a hotel that was very close to the beach.	1	2	3	4
	d) She got up to close the TV.	1	2	3	4
	e) There is always a close atmosphere in the countryside.	1	2	3	4
	f) I don't know				

8. Collocation recognition

In this final section of the test you need to choose the word that fits the best in the gap, the word that occurs more naturally in the context with the word underlined and in *italics*. All the options are possible to a greater or lesser degree, but there is only **one** best answer. Choose the most natural one, the best collocation of the word underlined and in *italics*.

Example:

	How sure are you of your answer?			
I left the building because I didn´t want to _____ any *trouble*. a) Make b) Get c) Create d) Cause e) I don´t know	1	2	3	4

Then, select how sure you are of your answers from 1 to 4 (**1** = not sure at all; **2** = not very sure; **3** = almost sure; **4** = totally sure). Please, **do not guess**. If you do not know the answer, just leave it blank.

		How sure are you of your answer?			
1.	She enjoys a lot of free time because she is only *employed* _____. a) Part-time b) Part-hours c) Short-time d) Some-days e) I don´t know	1	2	3	4

2.	The Government party did not *fulfill* their _____. a) Words b) Agreements c) Promises d) Speeches e) I don´t know	1	2	3	4
3.	He got a very high *mean* _____. a) Result b) Mark c) Number d) Score e) I don't know	1	2	3	4
4.	After his big mistake, my cousin tried very hard to *redeem* his _____. a) Popularity b) Respect c) Reputation d) Blame e) I don't know	1	2	3	4
5.	_____ *development* is important. a) Working b) Supportable c) Green d) Sustainable e) I don't know	1	2	3	4
6.	A text is easier to read if you *indent* the first _____. a) Line b) Word c) Number d) Letter e) I don't know	1	2	3	4

7.	It was *hard* _____. a) Money b) Work c) Traffic d) News e) I don't know	1	2	3	4
8.	She is the _____ *character* in the book. a) Important b) Chief c) Main d) Special e) I don't know	1	2	3	4
9.	Some people believe in _____ *charms*. a) Power b) Lucky c) Healthy d) Religious e) I don't know	1	2	3	4
10.	After a day in the office, I need to go outside for some *fresh* _____. a) Wind b) Sun c) Weather d) Air e) I don't know	1	2	3	4
11.	There are always more tourists in _____ *season*. a) Main b) Peak c) Big d) Top e) I don't know	1	2	3	4

12.	I like to drink *draught* _____ when it is hot in summer. a) Coke b) Juice c) Beer d) Cocktails e) I don't know	1	2	3	4
13.	They are *close* _____. a) Players b) Partners c) Workers d) Friends e) I don't know	1	2	3	4
14.	He has to _____ the *challenge*. a) Face b) Do c) Fill d) Connect e) I don't know	1	2	3	4
15.	You can *grate* the _____ into the bowl. a) Grapes b) Salt c) Cake d) Cheese e) I don´t know	1	2	3	4
16.	She has a *terminal* _____. a) Problem b) Disease c) Difficulty d) Sickness e) I don't know	1	2	3	4

17.	There was a *broad* _____ of colors. a) Range b) Number c) Mixture d) Amount e) I don't know	1	2	3	4
18.	In the summer people usually prefer *bright* _____. a) Beaches b) Colours c) Drinks d) Holidays e) I don´t know	1	2	3	4
19.	It is necessary to _____ a *distinction* between the two ideas. a) Create b) Make c) Do d) Perform e) I don't know	1	2	3	4
20.	I asked for a *bank* _____ to pay my many bills. a) Note b) Loan c) Notice d) Charge e) I don't know	1	2	3	4

In order to help us to better understand, interpret and classify your answers, would you mind telling us more about your personal and language learning background? Please provide the following information by ticking (✓) in the box or writing your response in the space.

Gender: ☐ Male ☐ Female

Age: _____

How many years have you been studying English in school/language centers? _____

Do you have a degree in English (e.g., Philology, Translation, Education)? ☐Yes ☐No

Are you studying English at the moment? ☐Yes ☐No **Where?** _____

Have you spent a long period (3 months or more) in English-speaking countries? ☐Yes ☐No

If your answer is Yes, how long? _____

Which is your level of English? ☐ Beginner ☐ Intermediate ☐ Advanced

How much time per week do you spend…:

- **reading books, magazines and newspapers in English, or visiting English language websites?** ☐ 0 hours ☐ 0-1 h. ☐ 1-2 h. ☐ 2+ h.
- **watching films, videos or TV in English?** ☐ 0 hours ☐ 0-1 h. ☐ 1-2 h. ☐ 2+ h.
- **listening to music in English?** ☐ 0 hours ☐ 0-1 h. ☐ 1-2 h. ☐ 2+ h.
- **using English to keep in contact with people? (Facebook, Twitter, Skype, email, SMS, etc.)**: ☐ 0 hours ☐ 0-1 h. ☐ 1-2 h. ☐ 2+ h.
- **playing computer/internet games in English**? ☐ 0 hours ☐ 0-1 h. ☐ 1-2 h. ☐ 2+ h.

Do you speak any other language, apart from English and your mother tongue? ☐Yes ☐No

Which one(s)? _____

This is the end of the test!

Congratulations for finishing it!

I would like to thank you very much for your cooperation! I appreciate your help and contribution to this survey a lot.

If you are interested in receiving information about the results of this survey, please feel free to email me:

beatriz.gonzalez-fernandez@nottingham.ac.uk

Thank you! ☺

Appendix 4. Participants' information sheet and sample of consent form.

The University of Nottingham

INFORMATION SHEET

As part of my PhD in the School of English, I am carrying out a study of vocabulary acquisition by learners of English as a second language. I am going to analyse how well different word aspects are learnt.

Because different aspects of words are tested, you will have to complete different sections. Each section will be given to you one by one, and you will have a time limit to complete it (this time limit is different depending on the section). At the beginning of each part you will have written instructions for completion, as well as some examples. Please, read them and ask any questions you may have. Then, answer the test trying not to guess. There will be breaks between sections so that you can rest.

You are free to withdraw from the study at any time. At every stage, your name will remain confidential. The data will be anonymized before the analysis and will be kept securely and used for academic purposes only.

Should you have any further queries about the study, please feel free to contact myself or my supervisor, Prof. Norbert Schmitt, who can be reached at norbert.schmitt@nottingham.ac.uk or by phone on +44 (0) 115 951 4847.

Signed

Beatriz González Fernández

beatriz.gonzalez-fernandez@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx



The University of Nottingham

HOJA INFORMATIVA

Como parte de mi doctorado, estoy llevando a cabo un estudio sobre la adquisición de vocabulario por estudiantes de inglés como segunda lengua. Para ello voy a analizar cómo y en qué medida se aprenden los distintos componentes de las palabras.

Debido a que varios aspectos de las palabras son examinados, hay que completar 8 secciones diferentes. Cada sección será distribuida una por una, y habrá un límite de tiempo para completarla (este límite de tiempo varía dependiendo de la sección). Al comienzo de cada parte hay instrucciones sobre cómo completar la sección, así como algunos ejemplos. Por favor, lee atentamente y pregunta cualquier duda que puedas tener. A continuación, responde a la prueba tratando de no adivinar lo que no sepas. Habrá pausas entre las secciones para que puedas descansar.

Eres libre de retirarte del estudio en cualquier momento. Tu nombre será tratado confidencialmente y los datos serán anónimos y se guardarán de forma segura, utilizándose solamente con fines académicos.

Si tienes cualquier preguntas sobre el estudio, no dudes en ponerse en contacto conmigo o con mi supervisor, el profesor Norbert Schmitt, en el email: norbert.schmitt@nottingham.ac.uk, o por teléfono al +44 (0) 115 951 4847.

Firmado

Beatriz González Fernández

beatriz.gonzalez-fernandez@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx

Informed Consent Form

Project title: **Investigating second language learners´ acquisition of vocabulary**

Please, read the following information and **circle** as necessary:

- I confirm that the purpose of the study has been explained and that I have understood it. (YES) / NO
- I have had the opportunity to ask questions and they have been successfully answered. (YES) / NO
- I understand that my participation in this study is voluntary and that I am free to withdraw from the study at any time, without having to give a reason and without consequence. (YES) / NO
- I understand that all data are anonymous and confidential, and that there will not be any connection between the personal information provided and the data. (YES) / NO
- I understand that there are no known risks or hazards associated with participating in this study. (YES) / NO
- I confirm that I have read and understood the attached information and that I agree to participate in this study. (YES) / NO
- I have received a copy of this Consent Form and of the accompanying Information Sheet. (YES) / NO

Participant's signature: _____ Date: 16/11/2015

Participant's Name (in block capitals):
ESTEFANÍA RODRÍGUEZ LUNOS

Researcher's signature: Beatriz Date: 16-11-15

Universidad de Nottingham
Facultad de inglés

Consentimiento Informado

Título del proyecto: **Investigación sobre la adquisición de vocabulario por estudiantes de segundas lenguas.**

Por favor, lea la siguiente información y rodee como corresponda:

- Confirmo que se me ha explicado el propósito de este estudio y que lo he entendido. SÍ / NO
- Confirmo que he tenido la oportunidad de hacer preguntas y éstas han sido respondidas adecuadamente. SÍ / NO
- Entiendo que la participación en este estudio es voluntaria y que soy libre de retirarme en cualquier momento sin tener que dar una razón y sin ninguna consecuencia. SÍ / NO
- Entiendo que los resultados son anónimos y confidenciales, y que no habrá ninguna relación entre los datos personales proporcionados y los datos del test. SÍ / NO
- Entiendo que no hay riesgos conocidos asociados a la participación en este estudio. SÍ / NO
- Confirmo que he leído y comprendido la información adjunta y que estoy de acuerdo en participar en este estudio. SÍ / NO
- He recibido una copia de este Consentimiento Informado y de la Hoja Informativa adjunta. SÍ / NO

Firma del participante: _____ Fecha: 17/11/15

Nombre del participante (en mayúsculas):
MARÍA PILAR SIERRA FERNÁNDEZ

Firma del investigador: _____ Fecha: 17-11-15

Appendix 5. Implicational Scaling matrix for the 80% accuracy criterion (frequencies) (Spanish).

Participant ID	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
21008	100	100	96.67	92.5	85	95	91.25	81.67
22057	100	85	91.67	96.25	90	95	91.25	83.33
22001	95	100	95.00	97.5	95	75	83.75	90.00
21010	100	100	85.00	88.75	95	90	88.75	80.00
21046	100	95	93.33	97.5	90	70	88.75	83.33
21016	90	100	93.33	78.75	90	95	80	86.67
21013	100	95	95.00	83.75	95	90	85	68.33
21061	95	95	90.00	86.25	85	90	83.75	78.33
21042	95	95	93.33	91.25	70	90	87.5	78.33
22056	95	100	80.00	93.75	75	75	91.25	80.00
22006	100	100	95.00	86.25	85	75	75	70.00
22005	95	90	91.67	83.75	95	75	75	76.67
22004	90	100	88.33	86.25	85	80	78.75	70.00
22058	100	100	85.00	86.25	85	75	76.25	68.33
21007	90	95	86.67	85	95	85	78.75	58.33
21048	95	95	88.33	85	85	80	80	63.33
21033	95	100	93.33	91.25	70	60	83.75	75.00
21015	100	90	78.33	86.25	80	80	73.75	73.33
22051	100	95	86.67	82.5	85	60	83.75	66.67
22059	90	95	73.33	87.5	85	80	80	63.33
21011	95	95	90.00	80	75	80	73.75	61.67
21064	100	100	78.33	86.25	90	80	66.25	48.33
21052	100	85	80.00	87.5	70	75	82.5	68.33
21006	100	95	78.33	82.5	80	70	78.75	56.67
22015	100	100	91.67	80	75	75	56.25	60.00
21050	90	100	90.00	78.75	70	75	65	68.33
22027	100	95	85.00	77.5	85	65	66.25	60.00
21040	95	95	78.33	71.25	80	80	71.25	61.67
22054	100	90	90.00	82.5	70	75	63.75	58.33
21012	95	85	88.33	85	70	70	76.25	60.00
22003	95	85	81.67	80	85	65	78.75	58.33
22034	100	90	83.33	90	80	55	72.5	55.00
22014	90	100	90.00	75	80	50	67.5	71.67
22041	95	90	86.67	78.75	70	70	68.75	61.67
21038	95	100	70.00	80	75	70	71.25	50.00
21018	90	90	90.00	87.5	65	60	75	53.33

22036	95	95	81.67	78.75	65	60	66.25	68.33
22017	100	90	78.33	75	80	65	57.5	63.33
21049	90	100	81.67	77.5	80	55	62.5	60.00
22024	85	85	86.67	81.25	70	55	73.75	66.67
22002	90	90	80.00	88.75	65	55	76.25	55.00
21060	90	90	81.67	82.5	60	65	70	58.33
21003	95	85	88.33	77.5	70	70	70	41.67
22040	85	85	73.33	75	80	75	68.75	55.00
21002	95	95	86.67	68.75	65	60	66.25	58.33
22016	85	95	75.00	75	80	55	71.25	58.33
22049	95	85	80.00	80	65	65	63.75	56.67
22045	90	95	81.67	66.25	80	60	61.25	53.33
22037	90	80	85.00	77.5	55	70	72.5	50.00
22011	90	85	80.00	75	65	60	66.25	58.33
21059	95	85	71.67	71.25	85	55	61.25	53.33
21026	85	85	76.67	67.5	70	75	57.5	58.33
21037	90	90	71.67	71.25	65	55	61.25	70.00
22035	90	100	76.67	56.25	70	60	61.25	56.67
22018	90	90	81.67	68.75	65	65	60	50.00
21009	85	90	71.67	63.75	80	70	52.5	51.67
21019	75	90	81.67	75	65	60	65	50.00
22022	80	95	90.00	68.75	65	45	61.25	56.67
22007	90	85	76.67	78.75	65	45	63.75	56.67
21001	85	90	66.67	72.5	90	45	57.5	50.00
12003	85	90	85.00	63.75	65	50	51.25	65.00
21004	85	85	90.00	52.5	75	60	47.5	60.00
22048	95	65	70.00	77.5	70	65	58.75	46.67
11001	95	85	83.33	60	70	50	42.5	58.33
21041	95	85	88.33	62.5	60	40	51.25	58.33
22019	85	100	71.67	62.5	60	45	60	43.33
22021	85	85	65.00	55	70	60	46.25	55.00
22047	95	90	60.00	63.75	75	50	38.75	48.33
21056	90	75	68.33	60	65	60	55	46.67
22012	85	85	73.33	58.75	60	55	51.25	46.67
22055	90	90	81.67	60	60	40	40	51.67
11002	90	75	76.67	57.5	55	55	50	53.33
21047	75	75	78.33	53.75	70	55	47.5	56.67
22046	90	75	76.67	58.75	50	55	47.5	55.00
22032	95	80	60.00	46.25	55	85	40	46.67
21021	80	80	70.00	46.25	60	65	46.25	56.67
21020	95	90	65.00	55	55	45	53.75	43.33
21063	75	90	63.33	57.5	70	60	50	35.00
21024	95	95	66.67	60	45	65	36.25	36.67
22020	80	95	55.00	63.75	65	50	53.75	36.67
21014	70	85	73.33	48.75	55	60	48.75	55.00
21005	85	95	75.00	51.25	65	35	40	45.00

21062	75	85	61.67	58.75	65	50	62.5	33.33
21065	90	85	66.67	55	55	40	50	46.67
22028	95	65	71.67	63.75	55	35	50	51.67
21017	75	90	71.67	58.75	70	35	37.5	48.33
21039	90	65	73.33	63.75	45	50	52.5	46.67
22009	80	90	80.00	56.25	55	40	38.75	43.33
21029	85	90	51.67	42.5	60	70	38.75	45.00
21058	80	65	63.33	63.75	60	55	48.75	45.00
21057	85	80	66.67	56.25	55	65	43.75	28.33
22008	90	80	53.33	61.25	45	55	48.75	45.00
22050	90	80	60.00	47.5	55	40	43.75	58.33
21031	80	85	65.00	43.75	70	40	35	50.00
22025	70	90	65.00	53.75	60	35	41.25	53.33
22013	85	85	63.33	46.25	40	60	37.5	45.00
22042	65	75	61.67	56.25	60	40	56.25	46.67
22038	90	55	73.33	48.75	45	50	41.25	40.00
21030	75	95	61.67	55	50	35	28.75	35.00
12009	80	75	68.33	33.75	50	40	43.75	41.67
21045	70	70	60.00	47.5	55	45	31.25	43.33
22053	90	65	66.67	50	30	45	40	35.00
21044	75	70	61.67	43.75	50	35	31.25	48.33
22043	70	60	68.33	43.75	60	40	36.25	35.00
21055	70	80	45.00	37.5	60	50	30	40.00
21043	75	80	58.33	35	60	40	25	38.33
11006	75	60	60.00	57.5	25	60	33.75	40.00
22044	80	70	63.33	48.75	35	30	38.75	36.67
21034	70	70	60.00	46.25	40	40	32.5	43.33
12008	70	70	56.67	55	35	30	42.5	38.33
21027	80	65	51.67	38.75	55	40	28.75	36.67
21023	75	65	53.33	53.75	40	30	40	36.67
11004	80	50	56.67	36.25	55	45	35	33.33
22010	60	65	65.00	41.25	50	30	28.75	41.67
22026	80	75	50.00	37.5	45	30	22.5	40.00
21054	70	55	53.33	48.75	35	50	32.5	31.67
22031	70	65	51.67	43.75	45	30	35	35.00
11003	65	60	73.33	31.25	35	45	22.5	41.67
21035	70	40	60.00	45	45	35	36.25	41.67
11010	65	65	56.67	41.25	40	45	25	35.00
12006	65	70	60.00	52.5	30	25	37.5	30.00
22029	60	75	61.67	37.5	40	25	36.25	33.33
12005	60	75	46.67	32.5	45	30	30	48.33
22033	65	70	65.00	48.75	45	25	26.25	21.67
22052	70	70	66.67	21.25	60	40	20	11.67
21051	55	60	53.33	50	30	30	43.75	33.33
21032	55	45	48.33	52.5	45	35	35	30.00
12004	80	40	65.00	30	35	30	18.75	38.33

22023	65	60	56.67	18.75	45	30	16.25	41.67
21053	70	45	55.00	48.75	20	25	35	33.33
22039	55	60	40.00	31.25	35	45	30	35.00
21028	65	65	45.00	27.5	40	35	22.5	26.67
21036	65	25	51.67	31.25	35	45	36.25	33.33
21025	65	50	38.33	45	35	40	12.5	31.67
12002	65	55	48.33	43.75	30	30	21.25	23.33
12001	55	65	61.67	33.75	35	15	17.5	28.33
11009	60	45	50.00	32.5	35	30	15	35.00
21022	60	65	43.33	20	25	20	28.75	36.67
12007	65	45	50.00	31.25	20	35	26.25	23.33
22030	65	55	43.33	27.5	40	25	5	21.67
11008	65	50	40.00	32.5	20	20	30	25.00
12010	45	40	35.00	30	40	35	22.5	10.00
11005	50	25	50.00	27.5	30	30	23.75	16.67
11007	40	40	31.67	25	15	10	6.25	6.67
TOTALS	11925	11475	10190	8795	8735	7720	7436.25	7155
Errors	20	15	19	10	17	10	5	1
Total 1s	97	91	50	34	32	16	15	7
Total 0s	47	53	94	110	112	128	129	137

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.92**, Maximum Marginals reproducibility = 0.78, Percentage of improvement in reproducibility= 0.14, Coefficient of scalability= **0.64.***

Appendix 6. Implicational Scaling matrix for the 80% accuracy criterion (Spanish).

Participant ID	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
21008	1	1	1	1	1	1	1	1
22057	1	1	1	1	1	1	1	1
21010	1	1	1	1	1	1	1	1
22001	1	1	1	1	1	0	1	1
21046	1	1	1	1	1	0	1	1
21016	1	1	1	0	1	1	1	1
21013	1	1	1	1	1	1	1	0
21061	1	1	1	1	1	1	1	0
21048	1	1	1	1	1	1	1	0
21042	1	1	1	1	0	1	1	0
22056	1	1	1	1	0	0	1	1
22004	1	1	1	1	1	1	0	0
21007	1	1	1	1	1	1	0	0
22051	1	1	1	1	1	0	1	0
22059	1	1	0	1	1	1	1	0
22006	1	1	1	1	1	0	0	0
22005	1	1	1	1	1	0	0	0
22058	1	1	1	1	1	0	0	0
21033	1	1	1	1	0	0	1	0
21015	1	1	0	1	1	1	0	0
21011	1	1	1	1	0	1	0	0
21064	1	1	0	1	1	1	0	0
21052	1	1	1	1	0	0	1	0
22003	1	1	1	1	1	0	0	0
22034	1	1	1	1	1	0	0	0
21006	1	1	0	1	1	0	0	0
22015	1	1	1	1	0	0	0	0
22027	1	1	1	0	1	0	0	0
21040	1	1	0	0	1	1	0	0
22054	1	1	1	1	0	0	0	0
21012	1	1	1	1	0	0	0	0
22014	1	1	1	0	1	0	0	0
21018	1	1	1	1	0	0	0	0

21049	1	1	1	0	1	0	0	0
22024	1	1	1	1	0	0	0	0
22002	1	1	1	1	0	0	0	0
21060	1	1	1	1	0	0	0	0
22049	1	1	1	1	0	0	0	0
22045	1	1	1	0	1	0	0	0
21050	1	1	1	0	0	0	0	0
22041	1	1	1	0	0	0	0	0
21038	1	1	0	1	0	0	0	0
22036	1	1	1	0	0	0	0	0
22017	1	1	0	0	1	0	0	0
21003	1	1	1	0	0	0	0	0
22040	1	1	0	0	1	0	0	0
21002	1	1	1	0	0	0	0	0
22016	1	1	0	0	1	0	0	0
22037	1	1	1	0	0	0	0	0
22011	1	1	1	0	0	0	0	0
21059	1	1	0	0	1	0	0	0
22018	1	1	1	0	0	0	0	0
21009	1	1	0	0	1	0	0	0
22022	1	1	1	0	0	0	0	0
21001	1	1	0	0	1	0	0	0
12003	1	1	1	0	0	0	0	0
21004	1	1	1	0	0	0	0	0
11001	1	1	1	0	0	0	0	0
21041	1	1	1	0	0	0	0	0
22055	1	1	1	0	0	0	0	0
22032	1	1	0	0	0	1	0	0
22009	1	1	1	0	0	0	0	0
21026	1	1	0	0	0	0	0	0
21037	1	1	0	0	0	0	0	0
22035	1	1	0	0	0	0	0	0
21019	0	1	1	0	0	0	0	0
22007	1	1	0	0	0	0	0	0
22019	1	1	0	0	0	0	0	0
22021	1	1	0	0	0	0	0	0
22047	1	1	0	0	0	0	0	0
22012	1	1	0	0	0	0	0	0
21021	1	1	0	0	0	0	0	0
21020	1	1	0	0	0	0	0	0
21024	1	1	0	0	0	0	0	0
22020	1	1	0	0	0	0	0	0

21005	1	1	0	0	0	0	0	0
21065	1	1	0	0	0	0	0	0
21029	1	1	0	0	0	0	0	0
21057	1	1	0	0	0	0	0	0
22008	1	1	0	0	0	0	0	0
22050	1	1	0	0	0	0	0	0
21031	1	1	0	0	0	0	0	0
22013	1	1	0	0	0	0	0	0
22048	1	0	0	0	0	0	0	0
21056	1	0	0	0	0	0	0	0
11002	1	0	0	0	0	0	0	0
22046	1	0	0	0	0	0	0	0
21063	0	1	0	0	0	0	0	0
21014	0	1	0	0	0	0	0	0
21062	0	1	0	0	0	0	0	0
22028	1	0	0	0	0	0	0	0
21017	0	1	0	0	0	0	0	0
21039	1	0	0	0	0	0	0	0
21058	1	0	0	0	0	0	0	0
22025	0	1	0	0	0	0	0	0
22038	1	0	0	0	0	0	0	0
21030	0	1	0	0	0	0	0	0
12009	1	0	0	0	0	0	0	0
22053	1	0	0	0	0	0	0	0
21055	0	1	0	0	0	0	0	0
21043	0	1	0	0	0	0	0	0
22044	1	0	0	0	0	0	0	0
21027	1	0	0	0	0	0	0	0
11004	1	0	0	0	0	0	0	0
22026	1	0	0	0	0	0	0	0
12004	1	0	0	0	0	0	0	0
21047	0	0	0	0	0	0	0	0
22042	0	0	0	0	0	0	0	0
21045	0	0	0	0	0	0	0	0
21044	0	0	0	0	0	0	0	0
22043	0	0	0	0	0	0	0	0
11006	0	0	0	0	0	0	0	0
21034	0	0	0	0	0	0	0	0
12008	0	0	0	0	0	0	0	0
21023	0	0	0	0	0	0	0	0
22010	0	0	0	0	0	0	0	0
21054	0	0	0	0	0	0	0	0

22031	0	0	0	0	0	0	0	0
11003	0	0	0	0	0	0	0	0
21035	0	0	0	0	0	0	0	0
11010	0	0	0	0	0	0	0	0
12006	0	0	0	0	0	0	0	0
22029	0	0	0	0	0	0	0	0
12005	0	0	0	0	0	0	0	0
22033	0	0	0	0	0	0	0	0
22052	0	0	0	0	0	0	0	0
21051	0	0	0	0	0	0	0	0
21032	0	0	0	0	0	0	0	0
22023	0	0	0	0	0	0	0	0
21053	0	0	0	0	0	0	0	0
22039	0	0	0	0	0	0	0	0
21028	0	0	0	0	0	0	0	0
21036	0	0	0	0	0	0	0	0
21025	0	0	0	0	0	0	0	0
12002	0	0	0	0	0	0	0	0
12001	0	0	0	0	0	0	0	0
11009	0	0	0	0	0	0	0	0
21022	0	0	0	0	0	0	0	0
12007	0	0	0	0	0	0	0	0
22030	0	0	0	0	0	0	0	0
11008	0	0	0	0	0	0	0	0
12010	0	0	0	0	0	0	0	0
11005	0	0	0	0	0	0	0	0
11007	0	0	0	0	0	0	0	0
Total 1s	97	91	50	34	32	16	15	7
Errors	9	10	16	8	16	10	4	1
Total 0s	47	53	94	110	112	128	129	137

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.94**, Maximum Marginals reproducibility = 0.78, Percentage of improvement in reproducibility= 0.16, Coefficient of scalability= **0.73.***

Appendix 7. Implicational Scaling matrix for the 75% accuracy criterion (frequencies) (Spanish).

Participant ID	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
21008	100	100	96.67	92.5	85	95	91.25	81.67
22057	100	85	91.67	96.25	90	95	91.25	83.33
22001	95	100	95.00	97.5	95	75	83.75	90.00
21010	100	100	85.00	88.75	95	90	88.75	80.00
21046	100	95	93.33	97.5	90	70	88.75	83.33
21016	90	100	93.33	78.75	90	95	80	86.67
21013	100	95	95.00	83.75	95	90	85	68.33
21061	95	95	90.00	86.25	85	90	83.75	78.33
21042	95	95	93.33	91.25	70	90	87.5	78.33
22056	95	100	80.00	93.75	75	75	91.25	80.00
22006	100	100	95.00	86.25	85	75	75	70.00
22005	95	90	91.67	83.75	95	75	75	76.67
22004	90	100	88.33	86.25	85	80	78.75	70.00
22058	100	100	85.00	86.25	85	75	76.25	68.33
21007	90	95	86.67	85	95	85	78.75	58.33
21048	95	95	88.33	85	85	80	80	63.33
21033	95	100	93.33	91.25	70	60	83.75	75.00
21015	100	90	78.33	86.25	80	80	73.75	73.33
22051	100	95	86.67	82.5	85	60	83.75	66.67
22059	90	95	73.33	87.5	85	80	80	63.33
21011	95	95	90.00	80	75	80	73.75	61.67
21064	100	100	78.33	86.25	90	80	66.25	48.33
21052	100	85	80.00	87.5	70	75	82.5	68.33
21006	100	95	78.33	82.5	80	70	78.75	56.67
22015	100	100	91.67	80	75	75	56.25	60.00
21050	90	100	90.00	78.75	70	75	65	68.33
22027	100	95	85.00	77.5	85	65	66.25	60.00
21040	95	95	78.33	71.25	80	80	71.25	61.67
22054	100	90	90.00	82.5	70	75	63.75	58.33
21012	95	85	88.33	85	70	70	76.25	60.00
22003	95	85	81.67	80	85	65	78.75	58.33
22034	100	90	83.33	90	80	55	72.5	55.00
22014	90	100	90.00	75	80	50	67.5	71.67
22041	95	90	86.67	78.75	70	70	68.75	61.67

21038	95	100	70.00	80	75	70	71.25	50.00
21018	90	90	90.00	87.5	65	60	75	53.33
22036	95	95	81.67	78.75	65	60	66.25	68.33
22017	100	90	78.33	75	80	65	57.5	63.33
21049	90	100	81.67	77.5	80	55	62.5	60.00
22024	85	85	86.67	81.25	70	55	73.75	66.67
22002	90	90	80.00	88.75	65	55	76.25	55.00
21060	90	90	81.67	82.5	60	65	70	58.33
21003	95	85	88.33	77.5	70	70	70	41.67
22040	85	85	73.33	75	80	75	68.75	55.00
21002	95	95	86.67	68.75	65	60	66.25	58.33
22016	85	95	75.00	75	80	55	71.25	58.33
22049	95	85	80.00	80	65	65	63.75	56.67
22045	90	95	81.67	66.25	80	60	61.25	53.33
22037	90	80	85.00	77.5	55	70	72.5	50.00
22011	90	85	80.00	75	65	60	66.25	58.33
21059	95	85	71.67	71.25	85	55	61.25	53.33
21026	85	85	76.67	67.5	70	75	57.5	58.33
21037	90	90	71.67	71.25	65	55	61.25	70.00
22035	90	100	76.67	56.25	70	60	61.25	56.67
22018	90	90	81.67	68.75	65	65	60	50.00
21009	85	90	71.67	63.75	80	70	52.5	51.67
21019	75	90	81.67	75	65	60	65	50.00
22022	80	95	90.00	68.75	65	45	61.25	56.67
22007	90	85	76.67	78.75	65	45	63.75	56.67
21001	85	90	66.67	72.5	90	45	57.5	50.00
12003	85	90	85.00	63.75	65	50	51.25	65.00
21004	85	85	90.00	52.5	75	60	47.5	60.00
22048	95	65	70.00	77.5	70	65	58.75	46.67
11001	95	85	83.33	60	70	50	42.5	58.33
21041	95	85	88.33	62.5	60	40	51.25	58.33
22019	85	100	71.67	62.5	60	45	60	43.33
22021	85	85	65.00	55	70	60	46.25	55.00
22047	95	90	60.00	63.75	75	50	38.75	48.33
21056	90	75	68.33	60	65	60	55	46.67
22012	85	85	73.33	58.75	60	55	51.25	46.67
22055	90	90	81.67	60	60	40	40	51.67
11002	90	75	76.67	57.5	55	55	50	53.33
21047	75	75	78.33	53.75	70	55	47.5	56.67
22046	90	75	76.67	58.75	50	55	47.5	55.00
22032	95	80	60.00	46.25	55	85	40	46.67
21021	80	80	70.00	46.25	60	65	46.25	56.67
21020	95	90	65.00	55	55	45	53.75	43.33
21063	75	90	63.33	57.5	70	60	50	35.00

21024	95	95	66.67	60	45	65	36.25	36.67
22020	80	95	55.00	63.75	65	50	53.75	36.67
21014	70	85	73.33	48.75	55	60	48.75	55.00
21005	85	95	75.00	51.25	65	35	40	45.00
21062	75	85	61.67	58.75	65	50	62.5	33.33
21065	90	85	66.67	55	55	40	50	46.67
22028	95	65	71.67	63.75	55	35	50	51.67
21017	75	90	71.67	58.75	70	35	37.5	48.33
21039	90	65	73.33	63.75	45	50	52.5	46.67
22009	80	90	80.00	56.25	55	40	38.75	43.33
21029	85	90	51.67	42.5	60	70	38.75	45.00
21058	80	65	63.33	63.75	60	55	48.75	45.00
21057	85	80	66.67	56.25	55	65	43.75	28.33
22008	90	80	53.33	61.25	45	55	48.75	45.00
22050	90	80	60.00	47.5	55	40	43.75	58.33
21031	80	85	65.00	43.75	70	40	35	50.00
22025	70	90	65.00	53.75	60	35	41.25	53.33
22013	85	85	63.33	46.25	40	60	37.5	45.00
22042	65	75	61.67	56.25	60	40	56.25	46.67
22038	90	55	73.33	48.75	45	50	41.25	40.00
21030	75	95	61.67	55	50	35	28.75	35.00
12009	80	75	68.33	33.75	50	40	43.75	41.67
21045	70	70	60.00	47.5	55	45	31.25	43.33
22053	90	65	66.67	50	30	45	40	35.00
21044	75	70	61.67	43.75	50	35	31.25	48.33
22043	70	60	68.33	43.75	60	40	36.25	35.00
21055	70	80	45.00	37.5	60	50	30	40.00
21043	75	80	58.33	35	60	40	25	38.33
11006	75	60	60.00	57.5	25	60	33.75	40.00
22044	80	70	63.33	48.75	35	30	38.75	36.67
21034	70	70	60.00	46.25	40	40	32.5	43.33
12008	70	70	56.67	55	35	30	42.5	38.33
21027	80	65	51.67	38.75	55	40	28.75	36.67
21023	75	65	53.33	53.75	40	30	40	36.67
11004	80	50	56.67	36.25	55	45	35	33.33
22010	60	65	65.00	41.25	50	30	28.75	41.67
22026	80	75	50.00	37.5	45	30	22.5	40.00
21054	70	55	53.33	48.75	35	50	32.5	31.67
22031	70	65	51.67	43.75	45	30	35	35.00
11003	65	60	73.33	31.25	35	45	22.5	41.67
21035	70	40	60.00	45	45	35	36.25	41.67
11010	65	65	56.67	41.25	40	45	25	35.00
12006	65	70	60.00	52.5	30	25	37.5	30.00
22029	60	75	61.67	37.5	40	25	36.25	33.33

12005	60	75	46.67	32.5	45	30	30	48.33
22033	65	70	65.00	48.75	45	25	26.25	21.67
22052	70	70	66.67	21.25	60	40	20	11.67
21051	55	60	53.33	50	30	30	43.75	33.33
21032	55	45	48.33	52.5	45	35	35	30.00
12004	80	40	65.00	30	35	30	18.75	38.33
22023	65	60	56.67	18.75	45	30	16.25	41.67
21053	70	45	55.00	48.75	20	25	35	33.33
22039	55	60	40.00	31.25	35	45	30	35.00
21028	65	65	45.00	27.5	40	35	22.5	26.67
21036	65	25	51.67	31.25	35	45	36.25	33.33
21025	65	50	38.33	45	35	40	12.5	31.67
12002	65	55	48.33	43.75	30	30	21.25	23.33
12001	55	65	61.67	33.75	35	15	17.5	28.33
11009	60	45	50.00	32.5	35	30	15	35.00
21022	60	65	43.33	20	25	20	28.75	36.67
12007	65	45	50.00	31.25	20	35	26.25	23.33
22030	65	55	43.33	27.5	40	25	5	21.67
11008	65	50	40.00	32.5	20	20	30	25.00
12010	45	40	35.00	30	40	35	22.5	10.00
11005	50	25	50.00	27.5	30	30	23.75	16.67
11007	40	40	31.67	25	15	10	6.25	6.67
TOTALS	11925	11475	10190	8795	8735	7720	7436.25	7155
Errors	10	12	15	6	17	8	7	3
Total 1s	107	100	63	50	38	27	25	11
Total 0s	37	44	81	94	106	117	119	133

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.93**, Maximum Marginals reproducibility = 0.74, Percentage of improvement in reproducibility= 0.19, Coefficient of scalability= **0.73.***

Appendix 8. Implicational Scaling matrix for the 75% accuracy criterion (Spanish).

Participant ID	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
21008	1	1	1	1	1	1	1	1
22057	1	1	1	1	1	1	1	1
22001	1	1	1	1	1	1	1	1
21010	1	1	1	1	1	1	1	1
21016	1	1	1	1	1	1	1	1
21061	1	1	1	1	1	1	1	1
22056	1	1	1	1	1	1	1	1
22005	1	1	1	1	1	1	1	1
21046	1	1	1	1	1	0	1	1
21013	1	1	1	1	1	1	1	0
21042	1	1	1	1	0	1	1	1
22006	1	1	1	1	1	1	1	0
22004	1	1	1	1	1	1	1	0
22058	1	1	1	1	1	1	1	0
21007	1	1	1	1	1	1	1	0
21048	1	1	1	1	1	1	1	0
21033	1	1	1	1	0	0	1	1
21015	1	1	1	1	1	1	0	0
22051	1	1	1	1	1	0	1	0
22059	1	1	0	1	1	1	1	0
21011	1	1	1	1	1	1	0	0
21064	1	1	1	1	1	1	0	0
21052	1	1	1	1	0	1	1	0
21006	1	1	1	1	1	0	1	0
22015	1	1	1	1	1	1	0	0
22003	1	1	1	1	1	0	1	0
21050	1	1	1	1	0	1	0	0
22027	1	1	1	1	1	0	0	0
21040	1	1	1	0	1	1	0	0
22054	1	1	1	1	0	1	0	0
21012	1	1	1	1	0	0	1	0
22034	1	1	1	1	1	0	0	0
22014	1	1	1	1	1	0	0	0

21018	1	1	1	1	0	0	1	0
22017	1	1	1	1	1	0	0	0
21049	1	1	1	1	1	0	0	0
22002	1	1	1	1	0	0	1	0
22040	1	1	0	1	1	1	0	0
22016	1	1	1	1	1	0	0	0
22041	1	1	1	1	0	0	0	0
21038	1	1	0	1	1	0	0	0
22036	1	1	1	1	0	0	0	0
22024	1	1	1	1	0	0	0	0
21060	1	1	1	1	0	0	0	0
21003	1	1	1	1	0	0	0	0
22049	1	1	1	1	0	0	0	0
22045	1	1	1	0	1	0	0	0
22037	1	1	1	1	0	0	0	0
22011	1	1	1	1	0	0	0	0
21026	1	1	1	0	0	1	0	0
21019	1	1	1	1	0	0	0	0
22007	1	1	1	1	0	0	0	0
21004	1	1	1	0	1	0	0	0
21002	1	1	1	0	0	0	0	0
21059	1	1	0	0	1	0	0	0
22035	1	1	1	0	0	0	0	0
22018	1	1	1	0	0	0	0	0
21009	1	1	0	0	1	0	0	0
22022	1	1	1	0	0	0	0	0
21001	1	1	0	0	1	0	0	0
12003	1	1	1	0	0	0	0	0
11001	1	1	1	0	0	0	0	0
21041	1	1	1	0	0	0	0	0
22047	1	1	0	0	1	0	0	0
22055	1	1	1	0	0	0	0	0
11002	1	1	1	0	0	0	0	0
21047	1	1	1	0	0	0	0	0
22046	1	1	1	0	0	0	0	0
22032	1	1	0	0	0	1	0	0
21005	1	1	1	0	0	0	0	0
22009	1	1	1	0	0	0	0	0
21037	1	1	0	0	0	0	0	0
22048	1	0	0	1	0	0	0	0
22019	1	1	0	0	0	0	0	0
22021	1	1	0	0	0	0	0	0

21056	1	1	0	0	0	0	0	0
22012	1	1	0	0	0	0	0	0
21021	1	1	0	0	0	0	0	0
21020	1	1	0	0	0	0	0	0
21063	1	1	0	0	0	0	0	0
21024	1	1	0	0	0	0	0	0
22020	1	1	0	0	0	0	0	0
21062	1	1	0	0	0	0	0	0
21065	1	1	0	0	0	0	0	0
21017	1	1	0	0	0	0	0	0
21029	1	1	0	0	0	0	0	0
21057	1	1	0	0	0	0	0	0
22008	1	1	0	0	0	0	0	0
22050	1	1	0	0	0	0	0	0
21031	1	1	0	0	0	0	0	0
22013	1	1	0	0	0	0	0	0
21030	1	1	0	0	0	0	0	0
12009	1	1	0	0	0	0	0	0
21043	1	1	0	0	0	0	0	0
22026	1	1	0	0	0	0	0	0
21014	0	1	0	0	0	0	0	0
22028	1	0	0	0	0	0	0	0
21039	1	0	0	0	0	0	0	0
21058	1	0	0	0	0	0	0	0
22025	0	1	0	0	0	0	0	0
22042	0	1	0	0	0	0	0	0
22038	1	0	0	0	0	0	0	0
22053	1	0	0	0	0	0	0	0
21044	1	0	0	0	0	0	0	0
21055	0	1	0	0	0	0	0	0
11006	1	0	0	0	0	0	0	0
22044	1	0	0	0	0	0	0	0
21027	1	0	0	0	0	0	0	0
21023	1	0	0	0	0	0	0	0
11004	1	0	0	0	0	0	0	0
22029	0	1	0	0	0	0	0	0
12005	0	1	0	0	0	0	0	0
12004	1	0	0	0	0	0	0	0
21045	0	0	0	0	0	0	0	0
22043	0	0	0	0	0	0	0	0
21034	0	0	0	0	0	0	0	0
12008	0	0	0	0	0	0	0	0

22010	0	0	0	0	0	0	0	0
21054	0	0	0	0	0	0	0	0
22031	0	0	0	0	0	0	0	0
11003	0	0	0	0	0	0	0	0
21035	0	0	0	0	0	0	0	0
11010	0	0	0	0	0	0	0	0
12006	0	0	0	0	0	0	0	0
22033	0	0	0	0	0	0	0	0
22052	0	0	0	0	0	0	0	0
21051	0	0	0	0	0	0	0	0
21032	0	0	0	0	0	0	0	0
22023	0	0	0	0	0	0	0	0
21053	0	0	0	0	0	0	0	0
22039	0	0	0	0	0	0	0	0
21028	0	0	0	0	0	0	0	0
21036	0	0	0	0	0	0	0	0
21025	0	0	0	0	0	0	0	0
12002	0	0	0	0	0	0	0	0
12001	0	0	0	0	0	0	0	0
11009	0	0	0	0	0	0	0	0
21022	0	0	0	0	0	0	0	0
12007	0	0	0	0	0	0	0	0
22030	0	0	0	0	0	0	0	0
11008	0	0	0	0	0	0	0	0
12010	0	0	0	0	0	0	0	0
11005	0	0	0	0	0	0	0	0
11007	0	0	0	0	0	0	0	0
Total 1s	107	100	63	50	38	27	25	11
Errors	6	9	8	4	15	9	7	3
Total 0s	37	44	81	94	106	117	119	133

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.95**, Maximum Marginals reproducibility = 0.74, Percentage of improvement in reproducibility= 0.21, Coefficient of scalability= **0.81.***

Appendix 9. Table of factor loadings, *z* scores and significance values for the Form–meaning Recall model with 20 target words as indicators.

Paths	**Factor Loadings R**	**R^2**	***Z***
Form–meaning Recall → mean	.630	.40	7.312***
Form–meaning Recall → hard	.486	.24	5.133***
Form–meaning Recall → employed	.337	.11	3.575***
Form–meaning Recall → fresh	.768	.60	10.482***
Form–meaning Recall → charm	.543	.30	5.648***
Form–meaning Recall → indent	.904	.82	19.355***
Form–meaning Recall → fulfil	.823	.68	12.929***
Form–meaning Recall → character	.174	.03	1.442 *p*=.15
Form–meaning Recall → redeem	.803	.65	11.253***
Form–meaning Recall →development	.716	.51	10.240***
Form–meaning Recall → close	.020	.00	0.246 *p*=.81
Form–meaning Recall → season	.632	.40	6.130***
Form–meaning Recall → bank	.213	.05	1.363 *p*=.17
Form–meaning Recall → terminal	.497	.25	4.948***
Form–meaning Recall → bright	.699	.49	9.569***
Form–meaning Recall → distinction	.509	.26	4.969***
Form–meaning Recall → draught	.748	.56	7.360***
Form–meaning Recall → challenge	.731	.53	11.223***
Form–meaning Recall → broad	.883	.78	14.976***
Form–meaning Recall → grates	.769	.59	10.209***

Appendix 10. Participants' information sheet and sample of consent form (Chinese).

The University of Nottingham

信息页

我是英语系的一名博士生，具体研究内容是将英语作为第二语言的学习者对词汇的掌握情况。因此我将对词汇各个方面的掌握情况逐一加以分析。

由于我将测验词汇不同方面的掌握情况，因此您需要完成与此相应的不同的部分。每部分的试题将依次分发给您，您需要在规定时间内做完（每部分的规定时间不同）。在每部分的开始您将读到协助您完成的书面说明与示例。请仔细阅读这些说明与示例，如果有疑问请务必向我提出。在答题中请不要猜答案。每个部分之间将有休息时间。

您有权利在测试中间的任何时间放弃实验。在实验的每一步您的姓名都将保密。在我们分析结果的过程中，所有的数据都会进行匿名处理并妥善保存，仅作研究使用。

如果您对此还有任何疑问，请联系我或我的导师 Norbert Schmitt 教授：邮箱 norbert.schmitt@nottingham.ac.uk，电话 +44 (0) 115 951 4847.

签名处：

Beatriz González Fernández

beatriz.gonzalez-fernandez@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx

University of Nottingham
School of English

知情同意表

项目标题：二语学习者词汇习得研究

请阅读以下信息，圈出符合您的情况：

- 我确认研究人员已向我解释研究目的，我对其也已完全理解。 (YES) / NO
- 我有机会向研究人员提出疑问，这些问题也得到了成功解答。 (YES) / NO
- 我了解我对这项研究的参与纯属自愿，并且可以在研究过程中任何时候退出，
- 退出时不许给出任何理由，退出后也不承担任何后果。 (YES) / NO
- 我明白所有的数据都是匿名并保密的，研究人员也不会在我提供的个人信息和
- 取得的数据之间作任何联系。 (YES) / NO
- 我明白参与本项研究不牵涉任何已知的危险。 (YES) / NO
- 我确认我已经阅读并理解附加信息，并确认同意参加本项研究。 (YES) / NO
- 我已获得一张知情同意表格与相应的信息页。 (YES) / NO

参与者签名: Peijue Lu 日期: 2016.11.5

参与者姓名（请用大写字母填写）：

PEIJUE LU

研究人员签名: [signature] 日期: 5-11-16

Appendix 11. Implicational Scaling matrix for the 80% accuracy criterion (frequencies) (Chinese).

Participant ID	FM Recog	Collo Recog	Deriv Recog	MM Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
11004	100	95	88.75	91.67	100	85	83.75	80.00
22010	100	100	85	91.67	95	80	63.75	73.33
12038	90	100	85	83.33	90	80	75	68.33
21002	95	85	83.75	90.00	75	80	73.75	68.33
11001	95	85	87.5	75.00	75	75	71.25	75.00
12025	95	85	81.25	81.67	75	75	70	71.67
12036	95	70	82.5	86.67	70	75	77.5	70.00
21001	90	85	86.25	80.00	65	75	67.5	71.67
21015	90	75	80	83.33	80	70	72.5	68.33
21014	90	90	83.75	85.00	80	65	58.75	63.33
11047	90	95	81.25	65.00	85	70	65	61.67
22002	90	85	87.5	88.33	65	60	70	66.67
12005	90	95	77.5	65.00	60	75	70	71.67
12021	100	95	80	81.67	60	70	62.5	55.00
12030	100	85	72.5	76.67	75	60	66.25	66.67
11045	85	90	82.5	80.00	65	55	65	75.00
11043	95	90	68.75	78.33	75	55	66.25	65.00
12006	90	85	80	71.67	75	60	66.25	61.67
12035	85	90	80	76.67	80	55	60	61.67
11034	80	85	77.5	76.67	70	60	66.25	66.67
11038	85	85	71.25	75.00	80	60	67.5	58.33
11066	85	90	77.5	73.33	65	70	61.25	58.33
21004	90	80	76.25	81.67	60	65	60	65.00
12017	85	90	75	81.67	70	60	51.25	61.67
12004	80	90	76.25	78.33	80	45	61.25	63.33
12041	95	85	75	71.67	85	50	58.75	53.33
11024	90	85	73.75	85.00	65	65	45	63.33
11036	85	90	72.5	70.00	70	65	63.75	53.33
22001	85	95	75	63.33	75	50	65	58.33
21011	85	90	62.5	65.00	75	75	58.75	55.00
12026	75	90	71.25	76.67	75	65	56.25	56.67
21012	90	85	76.25	71.67	65	55	58.75	63.33
12046	95	90	70	73.33	50	65	55	63.33
11011	85	85	75	68.33	60	65	60	60.00
21008	85	100	72.5	68.33	70	55	52.5	55.00
11028	80	85	58.75	73.33	75	60	60	60.00
12012	80	80	76.25	76.67	60	60	58.75	58.33
11041	80	85	57.5	76.67	80	60	55	51.67

12019	85	85	72.5	73.33	70	55	55	50.00
11030	75	90	80	75.00	60	50	60	55.00
12022	85	90	72.5	73.33	75	50	47.5	51.67
11026	85	75	71.25	81.67	60	55	61.25	55.00
11068	80	80	67.5	70.00	80	60	55	50.00
22007	90	85	62.5	75.00	80	60	45	45.00
12034	85	85	76.25	85.00	65	40	55	50.00
11072	75	85	73.75	65.00	55	60	73.75	53.33
12039	80	75	68.75	76.67	75	55	51.25	58.33
21006	75	85	67.5	66.67	70	65	55	53.33
22004	75	95	70	71.67	60	60	48.75	56.67
11021	90	75	65	76.67	55	65	55	55.00
11005	80	80	71.25	73.33	55	60	56.25	60.00
12029	80	90	70	66.67	75	40	60	53.33
12061	80	95	72.5	61.67	65	40	63.75	56.67
12052	75	85	52.5	80.00	80	60	53.75	48.33
22008	75	85	76.25	60.00	70	65	50	53.33
22011	85	95	62.5	65.00	65	60	51.25	48.33
11056	80	80	75	65.00	55	55	55	66.67
11049	90	85	67.5	76.67	55	50	60	46.67
11003	85	85	58.75	63.33	60	60	65	53.33
21013	85	85	73.75	63.33	55	60	56.25	50.00
12033	80	100	61.25	63.33	70	50	52.5	50.00
11052	80	75	70	66.67	60	70	51.25	53.33
12056	85	90	55	81.67	55	60	46.25	53.33
11007	80	85	70	63.33	75	45	51.25	55.00
11006	85	85	73.75	70.00	65	60	51.25	33.33
12023	85	85	75	65.00	60	40	63.75	48.33
12065	70	85	76.25	65.00	70	55	47.5	53.33
12002	80	90	68.75	61.67	65	55	52.5	46.67
12049	75	75	70	71.67	65	65	51.25	46.67
21003	80	85	66.25	60.00	75	60	41.25	51.67
22009	85	85	66.25	68.33	45	70	51.25	48.33
22005	80	85	65	70.00	60	60	42.5	55.00
22006	85	75	73.75	73.33	60	65	38.75	46.67
12047	80	80	63.75	70.00	65	55	47.5	55.00
11065	75	85	66.25	56.67	55	65	57.5	55.00
12016	90	70	77.5	66.67	60	45	55	50.00
12037	80	80	71.25	71.67	50	50	62.5	48.33
11013	85	80	68.75	73.33	40	65	47.5	53.33
21009	80	95	70	58.33	65	55	43.75	45.00
11018	80	90	61.25	55.00	85	45	58.75	36.67
12018	85	75	72.5	73.33	55	40	56.25	53.33
12014	85	85	65	73.33	45	45	51.25	56.67
11031	85	85	72.5	53.33	50	60	50	50.00
12059	90	70	66.25	73.33	45	60	51.25	50.00
12054	75	85	57.5	60.00	65	65	50	46.67
11017	85	85	65	68.33	60	35	56.25	48.33

11050	80	75	72.5	58.33	60	50	58.75	48.33
11035	80	75	66.25	58.33	45	60	58.75	58.33
11046	85	75	61.25	65.00	55	55	50	55.00
11012	85	80	57.5	68.33	65	45	50	48.33
11029	75	80	71.25	58.33	50	60	51.25	53.33
12008	75	75	71.25	66.67	55	55	53.75	46.67
12064	75	85	52.5	71.67	60	55	42.5	53.33
11010	90	70	75	65.00	75	30	43.75	45.00
11044	75	85	61.25	53.33	65	55	52.5	43.33
11002	95	80	51.25	63.33	60	45	42.5	51.67
21005	70	80	63.75	55.00	65	65	42.5	45.00
11061	75	80	60	56.67	55	55	57.5	45.00
12010	85	80	60	73.33	40	40	53.75	51.67
12050	75	70	72.5	60.00	60	50	51.25	45.00
12053	75	75	68.75	61.67	50	55	47.5	48.33
11051	85	70	60	71.67	70	40	46.25	36.67
11058	85	75	58.75	61.67	35	60	55	48.33
12072	85	70	73.75	60.00	50	50	45	45.00
12057	75	60	68.75	66.67	50	60	48.75	48.33
11032	80	60	65	56.67	50	60	53.75	51.67
11055	70	90	68.75	53.33	60	40	46.25	46.67
12020	75	60	70	58.33	60	45	55	51.67
12011	75	85	62.5	58.33	65	35	46.25	45.00
11057	75	80	62.5	55.00	50	55	48.75	45.00
12013	70	75	65	78.33	50	35	46.25	50.00
12051	80	65	50	63.33	50	60	47.5	53.33
12045	90	80	60	53.33	55	50	42.5	36.67
12031	75	75	57.5	55.00	60	55	41.25	48.33
11064	70	75	60	55.00	50	55	55	43.33
12071	80	65	67.5	61.67	35	50	56.25	45.00
11027	80	70	57.5	55.00	60	40	43.75	51.67
22003	70	75	43.75	61.67	55	60	41.25	50.00
21007	70	65	60	55.00	50	60	43.75	51.67
11019	65	70	68.75	68.33	40	40	53.75	48.33
12028	85	60	60	65.00	50	45	50	36.67
11053	70	65	61.25	58.33	50	45	52.5	48.33
11014	75	80	65	58.33	40	35	50	46.67
11016	60	75	66.25	66.67	50	35	48.75	48.33
12055	65	85	45	58.33	65	50	33.75	46.67
12001	85	75	68.75	50.00	45	30	52.5	41.67
11063	80	55	66.25	68.33	25	50	51.25	51.67
12003	75	70	67.5	45.00	35	55	56.25	43.33
12044	75	75	51.25	61.67	60	45	38.75	40.00
12063	75	80	60	46.67	45	50	45	45.00
11037	65	70	63.75	63.33	45	45	46.25	48.33
11054	70	45	70	71.67	45	50	51.25	41.67
11020	70	75	65	48.33	50	45	45	43.33
12043	65	80	58.75	53.33	55	45	42.5	40.00

11023	85	70	70	56.67	25	40	53.75	38.33
12040	75	65	66.25	55.00	40	50	48.75	36.67
12042	70	75	61.25	50.00	45	50	45	40.00
12058	90	60	63.75	68.33	30	45	33.75	43.33
12070	70	65	52.5	58.33	50	55	40	43.33
12015	75	90	57.5	53.33	45	40	46.25	26.67
11022	85	60	52.5	61.67	45	45	42.5	41.67
11067	80	45	60	51.67	40	70	46.25	36.67
11025	85	75	47.5	51.67	50	35	43.75	38.33
11062	70	80	53.75	68.33	20	45	42.5	46.67
11059	65	70	56.25	43.33	45	55	45	45.00
11069	75	55	55	60.00	40	55	40	43.33
21010	70	65	61.25	40.00	60	50	32.5	43.33
11042	70	70	43.75	50.00	45	50	46.25	46.67
12062	60	70	50	63.33	40	45	47.5	45.00
12032	65	60	72.5	48.33	60	45	38.75	30.00
11033	75	80	41.25	63.33	45	35	40	38.33
12027	70	55	60	55.00	45	40	47.5	43.33
12007	85	70	58.75	40.00	50	40	30	41.67
12067	70	70	51.25	46.67	30	55	45	40.00
11039	65	65	67.5	56.67	35	40	40	38.33
12066	65	60	47.5	56.67	50	50	37.5	38.33
11071	70	55	53.75	56.67	35	40	46.25	48.33
11040	75	50	56.25	51.67	35	50	40	38.33
11060	70	60	43.75	40.00	50	50	37.5	38.33
12024	70	70	48.75	46.67	45	35	33.75	40.00
11008	75	60	46.25	56.67	30	40	41.25	38.33
11048	60	55	65	56.67	35	40	37.5	36.67
12068	60	60	55	63.33	35	40	32.5	35.00
12060	60	60	56.25	51.67	35	40	35	36.67
12048	60	50	51.25	58.33	50	35	32.5	33.33
11015	70	55	47.5	60.00	25	35	40	25.00
12009	50	55	68.75	58.33	20	10	47.5	45.00
12069	75	55	37.5	38.33	40	30	33.75	35.00
11009	50	45	48.75	28.33	25	40	30	36.67
11070	65	40	27.5	33.33	35	30	28.75	41.67
TOTALS	13495	13145	11190	11033.3	9725	9035	8767.5	8550
Errors	33	27	4	11	15	0	0	0
Total 1s	99	94	16	19	15	4	1	1
Total 0s	71	76	154	151	155	166	169	169

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.93**, Maximum Marginals reproducibility = 0.85, Percentage of improvement in reproducibility= 0.08, Coefficient of scalability= **0.53.***

Appendix 12. Implicational Scaling matrix for the 80% accuracy criterion (Chinese).

Participant ID	FM Recog	Collo Recog	MM Recog	Deriv Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
11004	1	1	1	1	1	1	1	1
22010	1	1	1	1	1	1	0	0
12038	1	1	1	1	1	1	0	0
21002	1	1	1	1	0	1	0	0
21014	1	1	1	1	1	0	0	0
12025	1	1	1	1	0	0	0	0
21001	1	1	1	1	0	0	0	0
21015	1	0	1	1	1	0	0	0
11047	1	1	0	1	1	0	0	0
22002	1	1	1	1	0	0	0	0
12021	1	1	1	1	0	0	0	0
11045	1	1	1	1	0	0	0	0
12035	1	1	0	1	1	0	0	0
11001	1	1	0	1	0	0	0	0
12036	1	0	1	1	0	0	0	0
12006	1	1	0	1	0	0	0	0
11038	1	1	0	0	1	0	0	0
21004	1	1	1	0	0	0	0	0
12017	1	1	1	0	0	0	0	0
12004	1	1	0	0	1	0	0	0
12041	1	1	0	0	1	0	0	0
11024	1	1	1	0	0	0	0	0
11041	1	1	0	0	1	0	0	0
11068	1	1	0	0	1	0	0	0
22007	1	1	0	0	1	0	0	0
12034	1	1	1	0	0	0	0	0
12052	0	1	1	0	1	0	0	0
12056	1	1	1	0	0	0	0	0
11018	1	1	0	0	1	0	0	0
12005	1	1	0	0	0	0	0	0
12030	1	1	0	0	0	0	0	0
11043	1	1	0	0	0	0	0	0
11034	1	1	0	0	0	0	0	0
11066	1	1	0	0	0	0	0	0
11036	1	1	0	0	0	0	0	0
22001	1	1	0	0	0	0	0	0
21011	1	1	0	0	0	0	0	0
21012	1	1	0	0	0	0	0	0
12046	1	1	0	0	0	0	0	0

11011	1	1	0	0	0	0	0	0
21008	1	1	0	0	0	0	0	0
11028	1	1	0	0	0	0	0	0
12012	1	1	0	0	0	0	0	0
12019	1	1	0	0	0	0	0	0
11030	0	1	0	1	0	0	0	0
12022	1	1	0	0	0	0	0	0
11026	1	0	1	0	0	0	0	0
11005	1	1	0	0	0	0	0	0
12029	1	1	0	0	0	0	0	0
12061	1	1	0	0	0	0	0	0
22011	1	1	0	0	0	0	0	0
11056	1	1	0	0	0	0	0	0
11049	1	1	0	0	0	0	0	0
11003	1	1	0	0	0	0	0	0
21013	1	1	0	0	0	0	0	0
12033	1	1	0	0	0	0	0	0
11007	1	1	0	0	0	0	0	0
11006	1	1	0	0	0	0	0	0
12023	1	1	0	0	0	0	0	0
12002	1	1	0	0	0	0	0	0
21003	1	1	0	0	0	0	0	0
22009	1	1	0	0	0	0	0	0
22005	1	1	0	0	0	0	0	0
12047	1	1	0	0	0	0	0	0
12037	1	1	0	0	0	0	0	0
11013	1	1	0	0	0	0	0	0
21009	1	1	0	0	0	0	0	0
12014	1	1	0	0	0	0	0	0
11031	1	1	0	0	0	0	0	0
11017	1	1	0	0	0	0	0	0
11012	1	1	0	0	0	0	0	0
11002	1	1	0	0	0	0	0	0
12010	1	1	0	0	0	0	0	0
12045	1	1	0	0	0	0	0	0
12026	0	1	0	0	0	0	0	0
11072	0	1	0	0	0	0	0	0
12039	1	0	0	0	0	0	0	0
21006	0	1	0	0	0	0	0	0
22004	0	1	0	0	0	0	0	0
11021	1	0	0	0	0	0	0	0
22008	0	1	0	0	0	0	0	0
11052	1	0	0	0	0	0	0	0
12065	0	1	0	0	0	0	0	0
22006	1	0	0	0	0	0	0	0
11065	0	1	0	0	0	0	0	0
12016	1	0	0	0	0	0	0	0
12018	1	0	0	0	0	0	0	0

12059	1	0	0	0	0	0	0	0
12054	0	1	0	0	0	0	0	0
11050	1	0	0	0	0	0	0	0
11035	1	0	0	0	0	0	0	0
11046	1	0	0	0	0	0	0	0
11029	0	1	0	0	0	0	0	0
12064	0	1	0	0	0	0	0	0
11010	1	0	0	0	0	0	0	0
11044	0	1	0	0	0	0	0	0
21005	0	1	0	0	0	0	0	0
11061	0	1	0	0	0	0	0	0
11051	1	0	0	0	0	0	0	0
11058	1	0	0	0	0	0	0	0
12072	1	0	0	0	0	0	0	0
11032	1	0	0	0	0	0	0	0
11055	0	1	0	0	0	0	0	0
12011	0	1	0	0	0	0	0	0
11057	0	1	0	0	0	0	0	0
12051	1	0	0	0	0	0	0	0
12071	1	0	0	0	0	0	0	0
11027	1	0	0	0	0	0	0	0
12028	1	0	0	0	0	0	0	0
11014	0	1	0	0	0	0	0	0
12055	0	1	0	0	0	0	0	0
12001	1	0	0	0	0	0	0	0
11063	1	0	0	0	0	0	0	0
12063	0	1	0	0	0	0	0	0
12043	0	1	0	0	0	0	0	0
11023	1	0	0	0	0	0	0	0
12058	1	0	0	0	0	0	0	0
12015	0	1	0	0	0	0	0	0
11022	1	0	0	0	0	0	0	0
11067	1	0	0	0	0	0	0	0
11025	1	0	0	0	0	0	0	0
11062	0	1	0	0	0	0	0	0
11033	0	1	0	0	0	0	0	0
12007	1	0	0	0	0	0	0	0
12049	0	0	0	0	0	0	0	0
12008	0	0	0	0	0	0	0	0
12050	0	0	0	0	0	0	0	0
12053	0	0	0	0	0	0	0	0
12057	0	0	0	0	0	0	0	0
12020	0	0	0	0	0	0	0	0
12013	0	0	0	0	0	0	0	0
12031	0	0	0	0	0	0	0	0
11064	0	0	0	0	0	0	0	0
22003	0	0	0	0	0	0	0	0
21007	0	0	0	0	0	0	0	0

11019	0	0	0	0	0	0	0	0
11053	0	0	0	0	0	0	0	0
11016	0	0	0	0	0	0	0	0
12003	0	0	0	0	0	0	0	0
12044	0	0	0	0	0	0	0	0
11037	0	0	0	0	0	0	0	0
11054	0	0	0	0	0	0	0	0
11020	0	0	0	0	0	0	0	0
12040	0	0	0	0	0	0	0	0
12042	0	0	0	0	0	0	0	0
12070	0	0	0	0	0	0	0	0
11059	0	0	0	0	0	0	0	0
11069	0	0	0	0	0	0	0	0
21010	0	0	0	0	0	0	0	0
11042	0	0	0	0	0	0	0	0
12062	0	0	0	0	0	0	0	0
12032	0	0	0	0	0	0	0	0
12027	0	0	0	0	0	0	0	0
12067	0	0	0	0	0	0	0	0
11039	0	0	0	0	0	0	0	0
12066	0	0	0	0	0	0	0	0
11071	0	0	0	0	0	0	0	0
11040	0	0	0	0	0	0	0	0
11060	0	0	0	0	0	0	0	0
12024	0	0	0	0	0	0	0	0
11008	0	0	0	0	0	0	0	0
11048	0	0	0	0	0	0	0	0
12068	0	0	0	0	0	0	0	0
12060	0	0	0	0	0	0	0	0
12048	0	0	0	0	0	0	0	0
11015	0	0	0	0	0	0	0	0
12009	0	0	0	0	0	0	0	0
12069	0	0	0	0	0	0	0	0
11009	0	0	0	0	0	0	0	0
11070	0	0	0	0	0	0	0	0
Total 1s	99	94	19	17	15	4	1	1
Errors	25	25	10	1	12	0	0	0
Total 0s	71	76	151	153	155	166	169	169

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.95**, Maximum Marginals reproducibility = 0.85, Percentage of improvement in reproducibility= 0.10, Coefficient of scalability= **0.67.***

Appendix 13. Implicational Scaling matrix for the 75% accuracy criterion (frequencies) (Chinese).

Participant ID	FM Recog	Collo Recog	Deriv Recog	MM Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
11004	100	95	88.75	91.67	100	85	83.75	80.00
22010	100	100	85	91.67	95	80	63.75	73.33
12038	90	100	85	83.33	90	80	75	68.33
21002	95	85	83.75	90.00	75	80	73.75	68.33
11001	95	85	87.5	75.00	75	75	71.25	75.00
12025	95	85	81.25	81.67	75	75	70	71.67
12036	95	70	82.5	86.67	70	75	77.5	70.00
21001	90	85	86.25	80.00	65	75	67.5	71.67
21015	90	75	80	83.33	80	70	72.5	68.33
21014	90	90	83.75	85.00	80	65	58.75	63.33
11047	90	95	81.25	65.00	85	70	65	61.67
22002	90	85	87.5	88.33	65	60	70	66.67
12021	100	95	80	81.67	60	70	62.5	55.00
12005	90	95	77.5	65.00	60	75	70	71.67
12030	100	85	72.5	76.67	75	60	66.25	66.67
11045	85	90	82.5	80.00	65	55	65	75.00
11043	95	90	68.75	78.33	75	55	66.25	65.00
12006	90	85	80	71.67	75	60	66.25	61.67
12035	85	90	80	76.67	80	55	60	61.67
11034	80	85	77.5	76.67	70	60	66.25	66.67
11038	85	85	71.25	75.00	80	60	67.5	58.33
11066	85	90	77.5	73.33	65	70	61.25	58.33
21004	90	80	76.25	81.67	60	65	60	65.00
12017	85	90	75	81.67	70	60	51.25	61.67
12004	80	90	76.25	78.33	80	45	61.25	63.33
12041	95	85	75	71.67	85	50	58.75	53.33
11024	90	85	73.75	85.00	65	65	45	63.33
11036	85	90	72.5	70.00	70	65	63.75	53.33
22001	85	95	75	63.33	75	50	65	58.33
21011	85	90	62.5	65.00	75	75	58.75	55.00
12026	75	90	71.25	76.67	75	65	56.25	56.67
21012	90	85	76.25	71.67	65	55	58.75	63.33
12046	95	90	70	73.33	50	65	55	63.33
11011	85	85	75	68.33	60	65	60	60.00
21008	85	100	72.5	68.33	70	55	52.5	55.00
11028	80	85	58.75	73.33	75	60	60	60.00
12012	80	80	76.25	76.67	60	60	58.75	58.33
11041	80	85	57.5	76.67	80	60	55	51.67

12019	85	85	72.5	73.33	70	55	55	50.00
11030	75	90	80	75.00	60	50	60	55.00
12022	85	90	72.5	73.33	75	50	47.5	51.67
11026	85	75	71.25	81.67	60	55	61.25	55.00
11068	80	80	67.5	70.00	80	60	55	50.00
22007	90	85	62.5	75.00	80	60	45	45.00
12034	85	85	76.25	85.00	65	40	55	50.00
11072	75	85	73.75	65.00	55	60	73.75	53.33
12039	80	75	68.75	76.67	75	55	51.25	58.33
21006	75	85	67.5	66.67	70	65	55	53.33
22004	75	95	70	71.67	60	60	48.75	56.67
11021	90	75	65	76.67	55	65	55	55.00
11005	80	80	71.25	73.33	55	60	56.25	60.00
12029	80	90	70	66.67	75	40	60	53.33
12061	80	95	72.5	61.67	65	40	63.75	56.67
12052	75	85	52.5	80.00	80	60	53.75	48.33
22008	75	85	76.25	60.00	70	65	50	53.33
22011	85	95	62.5	65.00	65	60	51.25	48.33
11056	80	80	75	65.00	55	55	55	66.67
11049	90	85	67.5	76.67	55	50	60	46.67
11003	85	85	58.75	63.33	60	60	65	53.33
21013	85	85	73.75	63.33	55	60	56.25	50.00
12033	80	100	61.25	63.33	70	50	52.5	50.00
11052	80	75	70	66.67	60	70	51.25	53.33
12056	85	90	55	81.67	55	60	46.25	53.33
11007	80	85	70	63.33	75	45	51.25	55.00
11006	85	85	73.75	70.00	65	60	51.25	33.33
12023	85	85	75	65.00	60	40	63.75	48.33
12065	70	85	76.25	65.00	70	55	47.5	53.33
12002	80	90	68.75	61.67	65	55	52.5	46.67
12049	75	75	70	71.67	65	65	51.25	46.67
21003	80	85	66.25	60.00	75	60	41.25	51.67
22009	85	85	66.25	68.33	45	70	51.25	48.33
22005	80	85	65	70.00	60	60	42.5	55.00
22006	85	75	73.75	73.33	60	65	38.75	46.67
12047	80	80	63.75	70.00	65	55	47.5	55.00
11065	75	85	66.25	56.67	55	65	57.5	55.00
12016	90	70	77.5	66.67	60	45	55	50.00
12037	80	80	71.25	71.67	50	50	62.5	48.33
11013	85	80	68.75	73.33	40	65	47.5	53.33
21009	80	95	70	58.33	65	55	43.75	45.00
11018	80	90	61.25	55.00	85	45	58.75	36.67
12018	85	75	72.5	73.33	55	40	56.25	53.33
12014	85	85	65	73.33	45	45	51.25	56.67
11031	85	85	72.5	53.33	50	60	50	50.00
12059	90	70	66.25	73.33	45	60	51.25	50.00
12054	75	85	57.5	60.00	65	65	50	46.67
11017	85	85	65	68.33	60	35	56.25	48.33

11050	80	75	72.5	58.33	60	50	58.75	48.33
11035	80	75	66.25	58.33	45	60	58.75	58.33
11046	85	75	61.25	65.00	55	55	50	55.00
11012	85	80	57.5	68.33	65	45	50	48.33
11029	75	80	71.25	58.33	50	60	51.25	53.33
12008	75	75	71.25	66.67	55	55	53.75	46.67
12064	75	85	52.5	71.67	60	55	42.5	53.33
11010	90	70	75	65.00	75	30	43.75	45.00
11044	75	85	61.25	53.33	65	55	52.5	43.33
11002	95	80	51.25	63.33	60	45	42.5	51.67
21005	70	80	63.75	55.00	65	65	42.5	45.00
11061	75	80	60	56.67	55	55	57.5	45.00
12010	85	80	60	73.33	40	40	53.75	51.67
12050	75	70	72.5	60.00	60	50	51.25	45.00
12053	75	75	68.75	61.67	50	55	47.5	48.33
11051	85	70	60	71.67	70	40	46.25	36.67
11058	85	75	58.75	61.67	35	60	55	48.33
12072	85	70	73.75	60.00	50	50	45	45.00
12057	75	60	68.75	66.67	50	60	48.75	48.33
11032	80	60	65	56.67	50	60	53.75	51.67
11055	70	90	68.75	53.33	60	40	46.25	46.67
12020	75	60	70	58.33	60	45	55	51.67
12011	75	85	62.5	58.33	65	35	46.25	45.00
11057	75	80	62.5	55.00	50	55	48.75	45.00
12013	70	75	65	78.33	50	35	46.25	50.00
12051	80	65	50	63.33	50	60	47.5	53.33
12045	90	80	60	53.33	55	50	42.5	36.67
12031	75	75	57.5	55.00	60	55	41.25	48.33
11064	70	75	60	55.00	50	55	55	43.33
12071	80	65	67.5	61.67	35	50	56.25	45.00
11027	80	70	57.5	55.00	60	40	43.75	51.67
22003	70	75	43.75	61.67	55	60	41.25	50.00
21007	70	65	60	55.00	50	60	43.75	51.67
11019	65	70	68.75	68.33	40	40	53.75	48.33
12028	85	60	60	65.00	50	45	50	36.67
11053	70	65	61.25	58.33	50	45	52.5	48.33
11014	75	80	65	58.33	40	35	50	46.67
11016	60	75	66.25	66.67	50	35	48.75	48.33
12055	65	85	45	58.33	65	50	33.75	46.67
12001	85	75	68.75	50.00	45	30	52.5	41.67
11063	80	55	66.25	68.33	25	50	51.25	51.67
12003	75	70	67.5	45.00	35	55	56.25	43.33
12044	75	75	51.25	61.67	60	45	38.75	40.00
12063	75	80	60	46.67	45	50	45	45.00
11037	65	70	63.75	63.33	45	45	46.25	48.33
11054	70	45	70	71.67	45	50	51.25	41.67
11020	70	75	65	48.33	50	45	45	43.33
12043	65	80	58.75	53.33	55	45	42.5	40.00

11023	85	70	70	56.67	25	40	53.75	38.33
12040	75	65	66.25	55.00	40	50	48.75	36.67
12042	70	75	61.25	50.00	45	50	45	40.00
12070	70	65	52.5	58.33	50	55	40	43.33
12058	90	60	63.75	68.33	30	45	33.75	43.33
12015	75	90	57.5	53.33	45	40	46.25	26.67
11022	85	60	52.5	61.67	45	45	42.5	41.67
11067	80	45	60	51.67	40	70	46.25	36.67
11025	85	75	47.5	51.67	50	35	43.75	38.33
11062	70	80	53.75	68.33	20	45	42.5	46.67
11059	65	70	56.25	43.33	45	55	45	45.00
11069	75	55	55	60.00	40	55	40	43.33
21010	70	65	61.25	40.00	60	50	32.5	43.33
11042	70	70	43.75	50.00	45	50	46.25	46.67
12062	60	70	50	63.33	40	45	47.5	45.00
12032	65	60	72.5	48.33	60	45	38.75	30.00
11033	75	80	41.25	63.33	45	35	40	38.33
12027	70	55	60	55.00	45	40	47.5	43.33
12007	85	70	58.75	40.00	50	40	30	41.67
12067	70	70	51.25	46.67	30	55	45	40.00
11039	65	65	67.5	56.67	35	40	40	38.33
12066	65	60	47.5	56.67	50	50	37.5	38.33
11071	70	55	53.75	56.67	35	40	46.25	48.33
11040	75	50	56.25	51.67	35	50	40	38.33
11060	70	60	43.75	40.00	50	50	37.5	38.33
12024	70	70	48.75	46.67	45	35	33.75	40.00
11008	75	60	46.25	56.67	30	40	41.25	38.33
11048	60	55	65	56.67	35	40	37.5	36.67
12068	60	60	55	63.33	35	40	32.5	35.00
12060	60	60	56.25	51.67	35	40	35	36.67
12048	60	50	51.25	58.33	50	35	32.5	33.33
11015	70	55	47.5	60.00	25	35	40	25.00
12009	50	55	68.75	58.33	20	10	47.5	45.00
12069	75	55	37.5	38.33	40	30	33.75	35.00
11009	50	45	48.75	28.33	25	40	30	36.67
11070	65	40	27.5	33.33	35	30	28.75	41.67
TOTALS	13495	13145	11190	11033.3	9725	9035	8767.5	8550
Errors	24	27	16	18	24	2	2	2
Total 1s	132	118	35	35	31	10	3	3
Total 0s	38	52	135	135	139	160	167	167

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.92**, Maximum Marginals reproducibility = 0.85, Percentage of improvement in reproducibility= 0.07, Coefficient of scalability= **0.47.***

Appendix 14. Implicational Scaling matrix for the 75% accuracy criterion (Chinese).

Participant ID	FM Recog	Collo Recog	Deriv Recog	MM Recog	Collo Recall	FM Recall	Deriv Recall	MM Recall
11004	1	1	1	1	1	1	1	1
12038	1	1	1	1	1	1	1	0
11001	1	1	1	1	1	1	0	1
22010	1	1	1	1	1	1	0	0
21002	1	1	1	1	1	1	0	0
12025	1	1	1	1	1	1	0	0
12036	1	0	1	1	0	1	1	0
21001	1	1	1	1	0	1	0	0
21015	1	1	1	1	1	0	0	0
21014	1	1	1	1	1	0	0	0
11045	1	1	1	1	0	0	0	1
12035	1	1	1	1	1	0	0	0
12004	1	1	1	1	1	0	0	0
11047	1	1	1	0	1	0	0	0
22002	1	1	1	1	0	0	0	0
12021	1	1	1	1	0	0	0	0
12005	1	1	1	0	0	1	0	0
12030	1	1	0	1	1	0	0	0
11043	1	1	0	1	1	0	0	0
12006	1	1	1	0	1	0	0	0
11034	1	1	1	1	0	0	0	0
11038	1	1	0	1	1	0	0	0
21004	1	1	1	1	0	0	0	0
12017	1	1	1	1	0	0	0	0
12041	1	1	1	0	1	0	0	0
22001	1	1	1	0	1	0	0	0
21011	1	1	0	0	1	1	0	0
12026	1	1	0	1	1	0	0	0
12012	1	1	1	1	0	0	0	0
11041	1	1	0	1	1	0	0	0
11030	1	1	1	1	0	0	0	0
22007	1	1	0	1	1	0	0	0
12034	1	1	1	1	0	0	0	0
12039	1	1	0	1	1	0	0	0
12052	1	1	0	1	1	0	0	0
11066	1	1	1	0	0	0	0	0

11024	1	1	0	1	0	0	0	0
21012	1	1	1	0	0	0	0	0
11011	1	1	1	0	0	0	0	0
11028	1	1	0	0	1	0	0	0
12022	1	1	0	0	1	0	0	0
11026	1	1	0	1	0	0	0	0
11068	1	1	0	0	1	0	0	0
11021	1	1	0	1	0	0	0	0
12029	1	1	0	0	1	0	0	0
22008	1	1	1	0	0	0	0	0
11056	1	1	1	0	0	0	0	0
11049	1	1	0	1	0	0	0	0
12056	1	1	0	1	0	0	0	0
11007	1	1	0	0	1	0	0	0
12023	1	1	1	0	0	0	0	0
21003	1	1	0	0	1	0	0	0
11018	1	1	0	0	1	0	0	0
11010	1	0	1	0	1	0	0	0
11036	1	1	0	0	0	0	0	0
12046	1	1	0	0	0	0	0	0
21008	1	1	0	0	0	0	0	0
12019	1	1	0	0	0	0	0	0
11072	1	1	0	0	0	0	0	0
21006	1	1	0	0	0	0	0	0
22004	1	1	0	0	0	0	0	0
11005	1	1	0	0	0	0	0	0
12061	1	1	0	0	0	0	0	0
22011	1	1	0	0	0	0	0	0
11003	1	1	0	0	0	0	0	0
21013	1	1	0	0	0	0	0	0
12033	1	1	0	0	0	0	0	0
11052	1	1	0	0	0	0	0	0
11006	1	1	0	0	0	0	0	0
12065	0	1	1	0	0	0	0	0
12002	1	1	0	0	0	0	0	0
12049	1	1	0	0	0	0	0	0
22009	1	1	0	0	0	0	0	0
22005	1	1	0	0	0	0	0	0
22006	1	1	0	0	0	0	0	0
12047	1	1	0	0	0	0	0	0
11065	1	1	0	0	0	0	0	0
12016	1	0	1	0	0	0	0	0
12037	1	1	0	0	0	0	0	0
11013	1	1	0	0	0	0	0	0

21009	1	1	0	0	0	0	0	0
12018	1	1	0	0	0	0	0	0
12014	1	1	0	0	0	0	0	0
11031	1	1	0	0	0	0	0	0
12054	1	1	0	0	0	0	0	0
11017	1	1	0	0	0	0	0	0
11050	1	1	0	0	0	0	0	0
11035	1	1	0	0	0	0	0	0
11046	1	1	0	0	0	0	0	0
11012	1	1	0	0	0	0	0	0
11029	1	1	0	0	0	0	0	0
12008	1	1	0	0	0	0	0	0
12064	1	1	0	0	0	0	0	0
11044	1	1	0	0	0	0	0	0
11002	1	1	0	0	0	0	0	0
11061	1	1	0	0	0	0	0	0
12010	1	1	0	0	0	0	0	0
12053	1	1	0	0	0	0	0	0
11058	1	1	0	0	0	0	0	0
12011	1	1	0	0	0	0	0	0
11057	1	1	0	0	0	0	0	0
12013	0	1	0	1	0	0	0	0
12045	1	1	0	0	0	0	0	0
12031	1	1	0	0	0	0	0	0
11014	1	1	0	0	0	0	0	0
12001	1	1	0	0	0	0	0	0
12044	1	1	0	0	0	0	0	0
12063	1	1	0	0	0	0	0	0
12015	1	1	0	0	0	0	0	0
11025	1	1	0	0	0	0	0	0
11033	1	1	0	0	0	0	0	0
12059	1	0	0	0	0	0	0	0
21005	0	1	0	0	0	0	0	0
12050	1	0	0	0	0	0	0	0
11051	1	0	0	0	0	0	0	0
12072	1	0	0	0	0	0	0	0
12057	1	0	0	0	0	0	0	0
11032	1	0	0	0	0	0	0	0
11055	0	1	0	0	0	0	0	0
12020	1	0	0	0	0	0	0	0
12051	1	0	0	0	0	0	0	0
11064	0	1	0	0	0	0	0	0
12071	1	0	0	0	0	0	0	0
11027	1	0	0	0	0	0	0	0

22003	0	1	0	0	0	0	0	0
12028	1	0	0	0	0	0	0	0
11016	0	1	0	0	0	0	0	0
12055	0	1	0	0	0	0	0	0
11063	1	0	0	0	0	0	0	0
12003	1	0	0	0	0	0	0	0
11020	0	1	0	0	0	0	0	0
12043	0	1	0	0	0	0	0	0
11023	1	0	0	0	0	0	0	0
12040	1	0	0	0	0	0	0	0
12042	0	1	0	0	0	0	0	0
12058	1	0	0	0	0	0	0	0
11022	1	0	0	0	0	0	0	0
11067	1	0	0	0	0	0	0	0
11062	0	1	0	0	0	0	0	0
11069	1	0	0	0	0	0	0	0
12007	1	0	0	0	0	0	0	0
11040	1	0	0	0	0	0	0	0
11008	1	0	0	0	0	0	0	0
12069	1	0	0	0	0	0	0	0
21007	0	0	0	0	0	0	0	0
11019	0	0	0	0	0	0	0	0
11053	0	0	0	0	0	0	0	0
11037	0	0	0	0	0	0	0	0
11054	0	0	0	0	0	0	0	0
12070	0	0	0	0	0	0	0	0
11059	0	0	0	0	0	0	0	0
21010	0	0	0	0	0	0	0	0
11042	0	0	0	0	0	0	0	0
12062	0	0	0	0	0	0	0	0
12032	0	0	0	0	0	0	0	0
12027	0	0	0	0	0	0	0	0
12067	0	0	0	0	0	0	0	0
11039	0	0	0	0	0	0	0	0
12066	0	0	0	0	0	0	0	0
11071	0	0	0	0	0	0	0	0
11060	0	0	0	0	0	0	0	0
12024	0	0	0	0	0	0	0	0
11048	0	0	0	0	0	0	0	0
12068	0	0	0	0	0	0	0	0
12060	0	0	0	0	0	0	0	0
12048	0	0	0	0	0	0	0	0
11015	0	0	0	0	0	0	0	0
12009	0	0	0	0	0	0	0	0

11009	0	0	0	0	0	0	0	0
11070	0	0	0	0	0	0	0	0
Total 1s	132	118	35	35	31	10	3	3
Errors	12	13	16	12	20	2	1	2
Total 0s	38	52	135	135	139	160	167	167

Note: Red coloured numbers = errors; Salmon shading = implicational scale.

*Coefficient of reproducibility= **0.94**, Maximum Marginals reproducibility = 0.85, Percentage of improvement in reproducibility= 0.09, Coefficient of scalability= **0.60.***