



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Extramural English and academic verb knowledge:
A longitudinal study of Viennese students majoring
in English“

verfasst von / submitted by

Mag. Katharina Ghamarian-Krenn, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Doktorin der Philosophie (Dr.phil.)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

A 792 343

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

English and American Studies

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Ute Smit

Acknowledgements

As is true for most pieces of research, on this cover there should not only be the name of the researcher, but also the names of many unsung heroes, who have supported me on my journey. Therefore, I would like to dedicate this section to express my heartfelt thanks to all the heroes of my adventure.

First and foremost, I would like to thank my supervisor Univ.-Prof. Mag. Dr. Ute Smit, who gave me the chance to start this journey and proved to be a patient, loving and devoted “Doktormama”. She has always shown profound belief in me, my work and my abilities, even if I did not, and has provided me with invaluable advice throughout this project. It was also Ute, who suggested applying for the uni:docs fellowship programme for doctoral candidates, which enabled me to enjoy a three year scholarship at University of Vienna. Thank you very much for enabling me to focus exclusively on my research for three years.

I would like to extend my sincere thanks to my colleagues at work, Christina, Elli, Alex, Vanja, Tatjana, Verena and Theresa who showed me unwavering support and always lent me an ear. Moreover, I would like to especially highlight the support of Marlene Schwarz, who did not only help me to find my tremendous supervisor, but who offered me her advice throughout the whole research project.

Special thanks also go to Silvia Bauer-Marschallinger, who has accompanied me throughout my whole journey. Thank you for being my guide from my very first steps at the English Department until now. What has started as a work relationship has turned out to be a wonderful friendship! Thank you for taking the time to read a big part of this thesis and offering me with valuable suggestions! My gratitude also goes to Pia Resnik, who together with Silvia has always strengthened my back at work and has offered me professional as well as emotional support.

Furthermore, I also have the great pleasure to be part of the EduLing Research Group, who were a great forum for taking my first academic steps. Special thanks go to Thomas, Armin and Helen, whose advice and encouragement I have greatly benefited from. I would also like to express my appreciation to all lecturers who agreed to open their classrooms to me and to all students who volunteered to take part in this study.

I am also extremely grateful to Luke Plonsky and Igor Yanovich for their statistical advice and for being passionate discussion partners about a topic that is maybe not thrilling for everybody.

Next, I would like to thank my friends who did not only support me on an emotional level, always cheering me on, but who also invested considerable time in helping me with my data and filling in my piloting questionnaire. Thank you Magdalena, Ricky and Mathias for your support and for knowing how to distract me from my work.

There is another very special person in my life that I am extremely grateful for: I would like to thank my (now) husband Raffael for his endless support and encouragement to complete my thesis. I am extremely grateful for all the time you invested in listening to my ideas, your patience and your support with formatting and making this thesis visually appealing. This journey would not have been the same without you.

And last but not least a heartfelt thanks goes out to my parents, who have made me to who I am today by encouraging me to be curious from the very first day of my life. Thank you for always believing in me and for supporting me in following my dreams!

Abstract

Building on past research (e.g. Lee, 2019; Lee & Dressman, 2018; Lee & Sylvén, 2021; Sundqvist & Sylvén, 2014) which has already highlighted the importance of Extramural English for L2 vocabulary knowledge, this study is aiming at a close investigation of the knowledge of academic verbs and its relation to EE practices of 147 undergraduate students majoring in English. Following Nation's (2013) taxonomy of word knowledge the focus lies on Viennese English major students' form-meaning recognition, form-meaning recall, multiple meaning recognition and collocation recognition knowledge of academic verbs across frequency levels.

For this purpose four tests, one for each word knowledge component, were developed, containing the same academic verbs. The second construct researched, namely Extramural English, being defined as "English outside the walls" (Sundqvist & Sylvén, 2016, p. 6). Using a self-designed questionnaire, the survey elicits students' involvement in English outside their educational context. By following the students for one academic year and testing them at several occasions light can be shed on the longitudinal development of both constructs researched and their relation to each other.

While both academic vocabulary and EE behaviour at tertiary level have already been researched in other settings (e.g. Busby, 2020; Lee & Sylvén, 2021; Sundqvist & Sylvén, 2014), this study presents a rare longitudinal perspective, revealing interesting relationships between the four word knowledge components researched and a remarkably stable involvement in Extramural English activities over time. Moreover, using single and multiple linear regression models, the influence of Extramural English activities on the different aspects of academic verb knowledge is investigated, showing that particular types of Extramural English activities are significant predictors for academic verb knowledge.

Table of contents

List of Figures	viii
List of Tables	xii
List of abbreviations.....	xvii
1 Introduction	1
2 Vocabulary knowledge.....	2
2.1 Definition of a word	2
2.2 Multiword units.....	5
2.2.1 Collocations.....	10
2.3 Types of words	15
2.3.1 High-frequency vocabulary	15
2.3.2 Low-frequency vocabulary.....	17
2.3.3 Mid-frequency vocabulary	18
2.3.4 Technical vocabulary.....	18
2.3.5 Academic vocabulary	19
2.4 Word knowledge	22
2.4.1 Productive and receptive knowledge	22
2.4.2 Vocabulary breadth and depth.....	24
2.4.2.1 Component approach.....	24
2.4.2.2 Developmental approach	28
2.4.2.3 Metaphorical approach	30
2.4.2.4 Challenges of vocabulary breadth and depth	31
2.4.2.5 The ontogenesis model	32
3 Vocabulary testing	38
3.1 Testing theory	39
3.1.1 Construct.....	39
3.1.2 Reliability.....	40
3.1.3 Validity	42
3.2 Existent vocabulary tests.....	44

4	Literature review vocabulary knowledge	57
4.1	What we already know about form-meaning recognition knowledge.....	61
4.1.1	Primary education.....	64
4.1.2	Secondary education	65
4.1.3	Tertiary education.....	71
4.2	What we already know about form-meaning recall knowledge	76
4.2.1	Primary and secondary education	79
4.2.2	Tertiary education.....	82
4.3	What we already know about multiple meaning recognition knowledge	84
4.4	What we already know about collocation recognition knowledge	89
4.5	Potential relationships between word knowledge components researched.....	94
5	Language learning beyond the classroom	98
5.1	The concept of Extramural English (EE)	105
5.2	Comparison of findings	109
5.2.1	EE preceding English formal instruction	114
5.2.2	EE at primary school level	118
5.2.3	EE at secondary school level	120
5.2.4	EE at tertiary level	128
5.3	Challenges of EE and possible solutions	134
6	Study design.....	139
6.1	Research questions	140
6.2	Data collection & Instruments	143
6.2.1	Collection of data related to vocabulary knowledge:.....	143
6.2.2	The Extramural English questionnaire	147
6.3	Participants.....	156
7	Test development	159
7.1	Form-meaning Recall Test of Academic Verbs.....	160
7.1.1	Basic considerations.....	160
7.1.2	Item selection	162
7.1.3	Item development	163
7.1.4	Piloting	166

7.1.5	1 st data collection.....	169
7.1.6	2 nd and 3 rd data collection.....	173
7.2	Form-meaning Recognition Test of Academic Verbs.....	174
7.2.1	Item development	174
7.2.2	Piloting	176
7.2.3	1 st data collection.....	178
7.2.4	2 nd and 3 rd data collection.....	181
7.3	Collocation Recognition Test of Academic Verbs	181
7.3.1	Item selection	181
7.3.2	Item development	182
7.3.3	Piloting 1	187
7.3.4	Piloting 2	196
7.3.5	2 nd data collection.....	200
7.4	Multiple Meaning Recognition Test of Academic Verbs.....	201
7.4.1	Item selection	202
7.4.2	Item development	202
7.4.3	Piloting	204
7.4.4	2 nd data collection.....	208
8	Results on vocabulary tests	210
8.1	1st data collection (T1).....	211
8.2	2nd data collection (T2)	213
8.3	3rd data collection (T3)	215
8.4	Development of word knowledge components	216
8.4.1	Form-meaning Recognition Test of Academic Verbs	217
8.4.2	Form-meaning Recall Test of Academic Verbs	218
8.4.3	Multiple Meaning Recognition Test of Academic Verbs	219
8.4.4	Collocation Recognition Test of Academic Verbs	220
8.5	SEM Models.....	221
9	Results on Extramural English behaviour	230
9.1	Involvement in EE activities for T2	231
9.2	Grouping of variables	237

9.3	Differences in EE behaviour regarding biographical background variables	245
9.3.1	Gender	245
9.3.2	Age of onset	247
9.3.3	Highest educational degree	250
9.3.4	Experience abroad	252
9.3.5	Speaking opportunity in spare time	252
9.4	Attention to new words	255
9.5	Recording techniques.....	255
9.6	Word knowledge aspects.....	257
9.7	Involvement in EE activities T3.....	260
9.8	Differences in EE behaviour regarding biographical background variables	265
9.8.1	Gender	265
9.8.2	Age of onset	266
9.8.3	Highest educational degree	267
9.8.4	Experience abroad	268
9.8.5	Speaking opportunity.....	268
9.9	Changes of attention to new words.....	270
9.10	Changes of recording techniques	270
9.11	Changes of word knowledge aspects	270
10	Results on the relation between the main concepts.....	271
10.1	Impact of EE behaviour on vocabulary knowledge of academic verbs.....	271
10.1.1	Second data collection.....	272
10.1.1.1	Form-meaning recognition	272
10.1.1.2	Form-meaning recall.....	279
10.1.1.3	Collocation recognition	281
10.1.1.4	Multiple meaning recognition	283
10.1.2	Third data collection	285
10.1.2.1	Form meaning recognition	285
10.1.2.2	Form meaning recall	287
10.1.2.3	Collocation recognition	289
10.1.2.4	Multiple meaning recognition	290

10.1.3	Longitudinal aspect.....	291
10.2	Biographical background variables & word knowledge aspects considered in an EE setting.....	294
11	Discussion.....	298
11.1	Developing vocabulary tests for academic verb knowledge.....	299
11.2	Academic verb knowledge at tertiary level in Austria	302
11.3	Development of academic verb knowledge.....	303
11.4	The nature of vocabulary knowledge.....	308
11.5	Extramural English behaviour of Austrian English major students	314
11.6	The effect of biographical variables on EE and its stability over time	318
11.7	The effect of EE on academic verb knowledge	320
11.8	Implications for the educational context	324
12	Conclusion.....	326
	References	329
	Digital Appendix.....	373
	Deutsche Zusammenfassung	374

List of Figures

Figure 2.1: Model of phraseological approach to collocation combining Cowie (1994, 1998a, 1998b), Howarth (1998a, 1998b) and Benson et al. (1997)	11
Figure 2.2: Types of vocabulary	17
Figure 2.3: Beck, McKeown & Kucan's model (2002) taken from Roessingh (2016)	20
Figure 2.4: Developmental model by Dale (1965)	28
Figure 2.5: Word web (Aitchison, 2004, p. 94)	30
Figure 2.6: The lexical space (see Daller, Milton & Treffers-Daller, 2007, p. 8)	30
Figure 2.7: Hypothesized model by González-Fernández & Schmitt (2019)	31
Figure 2.8: Illustration of the concepts of fuzziness and optimum taken from Bordag, Gor & Opitz (2022, p. 187)	33
Figure 2.9: Illustration of the dimension of linguistic domains taken from Bordag, Gor & Opitz (2022, p. 187)	33
Figure 2.10: Illustration of the dimension of mappings taken from Bordag, Gor & Opitz (2022, p. 188)	34
Figure 2.11: Illustration of the dimension of networks taken from Bordag, Gor & Opitz (2022, p. 189)	35
Figure 2.12: Illustration of the ontogenesis model for the phonological and semantic domain taken from Bordag, Gor & Opitz (2022, p. 191)	36
Figure 2.13: Illustration of primary domain for two activities	38
Figure 3.1: Formula for Cronbach's coefficient α (Bachmann, 2004, p. 163)	41
Figure 3.2: Sample item of a Yes/No test format	47
Figure 3.3: Example of an PVST item	48
Figure 3.4: Example of an item of the Receptive polyseme vocabulary test	51
Figure 4.1: Distribution of publication year of form-meaning recognition studies	61
Figure 4.2: Distribution of testing method in receptive vocabulary size studies reviewed	61
Figure 4.3: Geographical spread of form-meaning recognition studies reviewed	62
Figure 4.4: Educational level of participants involved in form-meaning recognition studies reviewed	64
Figure 4.5: Distribution of publication year of form-meaning recall studies	77
Figure 4.6: Geographical spread of form-meaning recall studies	77
Figure 4.7: Distribution of testing method in form-meaning recall studies	78
Figure 4.8: Educational level of participants involved in form-meaning recall studies	79
Figure 4.9: Distribution of publication year of multiple meaning recognition studies reviewed	84
Figure 4.10: Distribution of testing method in multiple meaning recognition studies	84
Figure 4.11: Geographical spread of multiple meaning recognition studies	85
Figure 4.12: Distribution of publication year of collocation recognition studies	89
Figure 4.13: Distribution of testing method in collocation recognition studies	89

Figure 4.14: Geographical spread of collocation recognition studies	90
Figure 4.15: SEM model with best fit indices by González-Fernández & Schmitt (2019, p. 16)	95
Figure 5.1: Informal learning according to Schugurensky (2000).....	100
Figure 5.2: Adapted LBC model of Benson (2011) and Schwarz (2020)	100
Figure 5.3: Model of language learning in and beyond the classroom	103
Figure 5.4: Using the model of language learning in and beyond the classroom for learning profiles (e.g., Hannah).....	104
Figure 5.5: Model of L2 English learning (Sundqvist & Sylvén, 2016, p. 10)	106
Figure 5.6: Classification of Extramural learning situations on basis of Sundqvist & Sylvén (2016).....	107
Figure 5.7: Model of Extramural English.....	108
Figure 5.8: Number of empirical studies on EE from 1996 to 2022	110
Figure 5.9: Development of study design from 1996-2022.....	111
Figure 5.10: Methodological choices of empirical EE studies	112
Figure 5.11: Educational background of participants in empirical EE studies.....	113
Figure 5.12: Extramural English Worldmap	114
Figure 6.1: Main research gaps addressed	139
Figure 6.2: Sum of data collected and analysed	144
Figure 6.3: Data collection process.....	146
Figure 6.4: Visualization of drop-out rate.....	156
Figure 8.1: Data distribution of Form-meaning Recall and Recognition Test_1st data collection.....	212
Figure 8.2: Data distribution of Multiple Meaning and Collocation Recognition Test_2nd data collection.....	214
Figure 8.3: Data distribution of all four tests_3rd data collection	215
Figure 8.4: Development of form-meaning recognition knowledge.....	217
Figure 8.5: Development and outliers of Form-meaning Recognition Test of Academic Verbs	218
Figure 8.6: Development of Form-meaning Recall Test of Academic Verbs.....	218
Figure 8.7: Development of form-meaning recall knowledge.....	219
Figure 8.8: Development of multiple meaning recognition knowledge.....	220
Figure 8.9: Development of collocation recognition knowledge	220
Figure 8.10: Two competing models of vocabulary knowledge	224
Figure 8.11: Model 1 of Vocabulary knowledge_Second data collection	227
Figure 8.12: Model 2 of Vocabulary knowledge_Second data collection	228
Figure 9.1: EE activities in which at least 50% of the participants are involved in nearly always (N=153).....	231
Figure 9.2: Reasons for the use of subtitles_EE questionnaire_T2	232

Figure 9.3: EE activities in which at least 50% of the participants are involved in nearly always or often (N=153)	234
Figure 9.4: Word cloud for other EE activities named in EE questionnaire_T2	237
Figure 9.5: 2-factor solution for English online (EO)_EE questionnaire_T2	239
Figure 9.6: Boxplot and Scatterplot with interpolation line_gender & playing computer games.....	246
Figure 9.7: Scatterplots with interpolation line for medium effects_gender	246
Figure 9.8: Scatterplot with interpolation line_ age of onset & playing computer games offline and online.....	248
Figure 9.9: Scatterplot with interpolation line_age of onset & writing and reading SMS or WhatsApp messages.....	249
Figure 9.10: Scatterplot with interpolation line_age of onset & speaking English in face to face interactions	249
Figure 9.11: Scatterplots with interpolation lines_highest educational degree & written communication activities.....	251
Figure 9.12: Scatterplots with interpolation lines_experience abroad & reading SMS or WhatsApp messages & writing e-mails	252
Figure 9.13: Scatterplot with interpolation line_speaking opportunities & communication-oriented reading	253
Figure 9.14: Scatterplots with interpolation lines_speaking opportunities & written communication & Creative writing.....	254
Figure 9.15: Scatterplot with interpolation line_speaking opportunities & English speaking	254
Figure 9.16: Correlation matrix of recording techniques used in an EE setting_T2	257
Figure 9.17: Reasons for the use of subtitles_EE questionnaire_T3	262
Figure 9.18: Word cloud for other EE activities named in EE questionnaire_T3	264
Figure 9.19: Scatterplots with interpolation line for computational variables_gender.....	265
Figure 9.20: Scatterplots with interpolation line for computational variables_age of onset	266
Figure 9.21: Scatterplots with interpolation line for computational variables_highest educational degree	267
Figure 9.22: Scatterplots with interpolation line for computational variables_speaking opportunity	269
Figure 10.1: Scatterplots form-meaning recognition knowledge and potential EE predictors	274
Figure 10.2: Power analysis linear multiple regression with Gplot (Faul et al., 2007)	275
Figure 10.3: Scatterplot matrix for EE variables	277
Figure 10.4: Shared & unique explained variability of form-meaning recognition by Fiction-oriented reading & Creative writing_2 nd collection	279

Figure 10.5: Shared & unique explained variability of collocation recognition by Creative writing and English speaking_2 nd collection	282
Figure 10.6: Development of total variance explained by EE predictor variables in a general population.....	293
Figure 11.1: Developmental curve form-meaning recognition knowledge	305
Figure 11.2: Developmental curves and growth rates form-meaning recognition and recall	306
Figure 11.3: Developmental curves dependent on frequency of use	307
Figure 11.4: Developmental curves for all word knowledge components measured	308
Figure 11.5: Illustration of the dimension of mappings taken from Bordag, Gor & Opitz (2022, p. 188).....	310
Figure 11.6: Development and Validation Procedures of the EE Scale Sundqvist & Uztosun (under review).....	315
Figure 11.7: Constructs of Sundqvist & Uztosun (under review) with reliability value	315

List of Tables

Table 2.1: Types of multiword units	7
Table 2.2: Form-based and meaning-based factors affecting multiword unit selection (Nation, 2013, p. 488)	14
Table 2.3: Nation's (2013) taxonomy of word knowledge	25
Table 4.1: Summary of studies considered in the literature review on vocabulary knowledge	60
Table 4.2: Relationships between word knowledge components reported in the literature	96
Table 5.1: Empirical studies on EE sorted by publication year	110
Table 5.2: Complementary functions of in- and out-of-school language learning.....	139
Table 6.1: Constructs of the EE questionnaire.....	148
Table 6.2: Reliability of constructs of EE questionnaire before and after piloting.....	149
Table 6.3: Link EE items with Nation's (2013) taxonomy	154
Table 6.4: Biographical profile of participant group.....	158
Table 7.1: Dimensions of Vocabulary assessment (Read, 2000, p. 9)	161
Table 7.2: Item selection 1st draft	163
Table 7.3: Lexical profile of Recall Test of Academic Verbs.....	164
Table 7.4: Descriptive results on Form-meaning Recall Test of Academic Verbs (N=36)	167
Table 7.5: Descriptive results for Form-meaning Recall Test of Academic Verbs questionnaire (N=34)	168
Table 7.6: Descriptive results for Form-meaning Recall Test of Academic Verbs 1st data collection (N=225).....	170
Table 7.7: Distribution of facility values of Form-meaning Recall Test of Academic Verbs (N=225)	171
Table 7.8: Categorization of discrimination values (Ebel, 1979)	171
Table 7.9: Discrimination values Form-meaning Recall Test of Academic Verbs (N=225)....	172
Table 7.10: Words outside the 2000 frequency band Form-meaning Recognition Test of Academic Verbs	175
Table 7.11: Descriptive results for Form-meaning Recognition Test of Academic Verbs (N=36)	177
Table 7.12: Descriptive results for Form-meaning Recognition Test of Academic Verbs questionnaire (N=34).....	178
Table 7.13: Descriptive results for Form-meaning Recognition Test of Academic Verbs 1st data collection (N=225).....	179
Table 7.14: Distribution of facility values of Form-meaning Recognition Test of Academic Verbs (N=225)	179
Table 7.15: Discrimination values Form-meaning Recognition Test of Academic Verbs (N=225)	180

Table 7.16: Discrimination values Form-meaning Recognition Test of Academic Verbs (N=225)	180
Table 7.17: Item selection for Collocation Recognition Test of Academic Verbs.....	182
Table 7.18: Lexical profile of Collocation Recognition Test of Academic Verbs.....	184
Table 7.19: Distribution of points for different scoring methods of WAT	185
Table 7.20: Descriptive results COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)	189
Table 7.21: Distribution of facility values COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)	189
Table 7.22: Discrimination values COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)	190
Table 7.23: Item facility with traditional one-point scoring COLLEX Collocation Recognition Test piloting 1 (N=19).....	191
Table 7.24: Discrimination values with traditional one-point scoring COLLEX Collocation Test piloting 1 (N=19)	192
Table 7.25: Descriptive results WAF format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)	192
Table 7.26: Item facility with partial one-point scoring WAF format piloting 1 (N=19).....	193
Table 7.27: Discrimination values partial one-point scoring WAF format piloting 1 (N=19)	193
Table 7.28: Item facility with traditional one-point scoring collocation WAF format piloting 1 (N=19)	194
Table 7.29: Discrimination values with traditional one-point scoring WAF format piloting 1 (N=19)	194
Table 7.30: Guessing behavior sorted by proficiency Collocation Recognition Test of Academic Verbs piloting 1 (N=16)	196
Table 7.31: Descriptive results COLLEX format piloting 2 (N=28)	197
Table 7.32: Item facility COLLEX format piloting 2 (N=28)	197
Table 7.33: Discrimination values COLLEX format piloting 2 (N=28).....	198
Table 7.34: Descriptive results WAF format piloting 2 (N=28)	199
Table 7.35: Item facility WAF format piloting 2 (N=28).....	199
Table 7.36: Descriptive results Collocation Recognition Test of Academic Verbs 2nd data collection (N=202).....	200
Table 7.37: Item facility Collocation Recognition Test of Academic Verbs 2nd data collection (N=202)	201
Table 7.38: Lexical profile of Multiple Meaning Recognition Test of Academic Verbs	204
Table 7.39: Descriptive results headword scoring Multiple Meaning Test (N=23)	205
Table 7.40: Item facility headword scoring Multiple Meaning Test (N=23)	205
Table 7.41: Discrimination values headword scoring Multiple Meaning Test piloting	206
Table 7.42: Relationship facility value, discrimination value and Cronbach alpha	206

Table 7.43: Descriptive statistics partial credit Multiple Meaning Test (N=23)	207
Table 7.44: Item facility partial credit Multiple Meaning Recognition Test (N=23)	207
Table 7.45: Discrimination values partial credit Multiple Meaning Recognition Test (N=23)	208
Table 7.46: Descriptive results for the Multiple Meaning Recognition Test, second data collection (N=204).....	209
Table 7.47: Item facility Multiple Meaning Recognition Test, 2nd data collection (N=202).210	
Table 7.48: Discrimination values Multiple Meaning Recognition Test, 2nd piloting (N=202)	210
Table 8.1: Descriptive statistics_1st data collection.....	211
Table 8.2: Correlations between form-meaning recognition and recall_1st data collection	212
Table 8.3: Descriptive statistics_2nd data collection	213
Table 8.4: Correlations between word knowledge components_2nd data collection	214
Table 8.5: Descriptive statistics_3rd data collection	215
Table 8.6: Correlations between word knowledge components_3rd data collection	216
Table 8.7: Selection of fit indices for SEM Model	222
Table 8.8: Model-fit indices for SEM Model 1 & 2_Second data collection	225
Table 8.9: Model-fit indices for SEM Model 1 & 2_Third data collection	225
Table 8.10: Summary of standardized regression coefficients for Model 1 for the second and third data collection.....	226
Table 8.11: Summary of standardized regression coefficients for Model 2 for the second and third data collection.....	229
Table 9.1: EE activities sorted by central tendency and interquartile range	236
Table 9.2: EE_Questionnaire subconstructs of Extramural behaviour.....	238
Table 9.3: Loading plot 4-factor solution Reading in English (RIE)_EE questionnaire_T2.....	240
Table 9.4: Loading plot 3-factor solution Writing in English (WIE)_EE questionnaire_T2	241
Table 9.5: Information about computational variables EE_questionnaire	244
Table 9.6: Significant differences found for computational variables_gender	245
Table 9.7: Grouping of participants by age of onset	247
Table 9.8: Significant differences found for computational variables_age of onset	247
Table 9.9: Significant differences found for computational variables_highest educational degree	250
Table 9.10: Significant differences found for computational variables_speaking opportunity	253
Table 9.11: Frequency of consideration of word knowledge aspects in free time sorted by median	259
Table 10.1: Correlation coefficients form-meaning recognition knowledge and potential EE predictors.....	273

Table 10.2: Model summary linear regression model form-meaning recognition, Fiction-oriented reading & Creative writing	276
Table 10.3: Model parameters form-meaning recognition & EE predictors_second data collection.....	276
Table 10.4: Model summary linear regression model form-meaning recognition & Fiction-oriented reading	278
Table 10.5: Model parameters form-meaning recognition & Fiction-oriented reading_2 nd data collection.....	278
Table 10.6: Correlation coefficients form-meaning recall and potential EE predictors.....	279
Table 10.7: Model summary linear regression model form-meaning recall & Creative writing	280
Table 10.8: Model parameters form-meaning recall & Creative writing_2 nd data collection	280
Table 10.9: Correlation coefficients collocation recognition and potential EE predictors....	281
Table 10.10: Model summary linear regression model collocation recognition & Creative writing	281
Table 10.11: Model summary linear regression model collocation recognition & English speaking	282
Table 10.12: Model parameters collocation recognition & EE predictors_2 nd data collection	282
Table 10.13: Correlation coefficients multiple meaning recognition and potential EE predictors.....	283
Table 10.14: Model summary linear regression model multiple meaning recognition & EE predictors.....	283
Table 10.15: Model parameters multiple meaning recognition & EE predictors_2 nd collection	284
Table 10.16: Model parameters multiple meaning recognition & Fiction-oriented reading_2 nd data collection.....	284
Table 10.17: Model parameters multiple meaning recognition & English online_2 nd data collection.....	285
Table 10.18: Correlation coefficients form-meaning recognition and potential EE predictors_3rd data collection.....	286
Table 10.19: Model summary linear regression form-meaning recognition & EE predictors_3rd data collection.....	286
Table 10.20: Model parameters form-meaning recognition & EE predictors_3rd data collection.....	287
Table 10.21: Correlation coefficients form-meaning recall and potential EE predictors_3rd data collection.....	287

Table 10.22: Model parameters form-meaning recall & Fiction-oriented reading_3rd data collection.....	287
Table 10.23: Model parameters form-meaning recall & Creative writing_3rd data collection	288
Table 10.24: Model parameters form-meaning recall & EE predictors_3rd data collection	288
Table 10.25: Correlation coefficients collocation recognition and potential EE predictors_3rd data collection.....	289
Table 10.26: Model parameters collocation recognition & English online_3rd data collection	289
Table 10.27: Model parameters collocation recognition & Fiction-oriented reading_3rd data collection.....	289
Table 10.28: Model parameters collocation recognition & EE predictors_3rd data collection	290
Table 10.29: Correlation coefficients multiple meaning recognition and potential EE predictors_3rd data collection.....	290
Table 10.30: Model parameters multiple meaning recognition & EE predictors_3rd data collection.....	291
Table 10.31: Model summary linear regression multiple meaning recognition & EE predictors_3rd data collection.....	291
Table 10.32: Summary of linear regression results for EE and all vocabulary tests.....	292
Table 10.33: Variability explained in the population observed.....	293
Table 10.34: Spearman correlation matrix biographical background variables & word knowledge components.....	294
Table 10.35: Spearman correlation matrix for consideration of word knowledge aspects for words encountered in EE activities.....	296
Table 10.36: Model parameters multiple meaning recognition & Word knowledge aspects predictors.....	297
Table 10.37: Model summary linear regression multiple meaning recognition & Word knowledge aspects predictors	297
Table 10.38: : Model parameters multiple meaning recognition & all predictors.....	298
Table 11.1: Correlation coefficients for all word knowledge components	310
Table 11.2: EE activities predicting four different word knowledge components of academic verbs.....	321
Table 11.3: Predictive power of EE activities.....	323
Table 11.4: Developmental patterns of predictive power	324

List of abbreviations

AVL.....	Academic Vocabulary List
AWL.....	Academic Word List
BE06.....	Corpus of British English and recent language change
BNC.....	British National Corpus
CATSS.....	Computer Adaptive Test of Size and Strength
CEFR.....	Common European Framework of Reference
CLIL.....	Content and language integrated learning
COCA.....	Corpus of Contemporary American English
COLLEX.....	Collocating lexis
DISCO.....	Discriminating Collocation Test
DVK.....	Depth of Vocabulary Knowledge Measure
EAP.....	English for academic purposes
EE.....	Extramural English
EFL.....	English as a foreign language
ELF.....	English as a lingua franca
EGAP.....	English for general academic purposes
EMI.....	English medium instruction
EnTenTen12.....	English Web Corpus
ESAP.....	English for special academic purposes
GSL.....	General Service List
ILSS1.....	Integrated Language and Study Skills 1
LOB.....	Lanchester-Oxford-Bergen Corpus
L1.....	First language/mother tongue
L2.....	Second language
LBC.....	Learning and teaching beyond the classroom
MANOVA.....	Multivariate analysis of variance
MLM.....	maximum likelihood method
NAWL.....	New Academic Word List
P.....	productive
PHRASE	Phrasal Vocabulary Size Test
PVLT.....	Productive Vocabulary Levels Test
PVST.....	Picture Vocabulary Size Test
R.....	receptive
SEM.....	Structural equation modelling
VELT.....	Vienna English Language Test
VLT.....	Vocabulary Levels Test
VKS.....	Vocabulary Knowledge Scale
VST.....	Vocabulary Size Test

WAT.....	Word Associates Test
YLVAT.....	Young Learners Vocabulary Assessment Test

Statistical abbreviations

α	Cronbach alpha
AGFI.....	adjusted goodness-of fit
B.....	unstandardized regression coefficient
β	standardized regression coefficient
CFI.....	comparative fit index
CI.....	confidence intervall
D	test statistic of the Kolmogorov-Smirnov test (normal distribution)
df	degrees of freedom
ϵ	epsilon
F	F-statistic
F	test statistic for Jonckheere-Terpstra test
GFI.....	goodness-of fit
H	test statistic of Kruskal-Wallis test
IQR.....	Interquartile Range
KMO.....	Kaiser-Meyer-Olkin measure of sampling adquacy
M	mean
Max.....	maximum
Mdn	median
Min.....	minimum
N	number of participants
η^2	eta squared effect size
p	p-value of statistical significance
r	Pearson's correlation coefficient
r^2	effect size of Spearman correlation
r	effect size estimate for Mann-Whitney test
RMSEA.....	root mean squared error of approximation
R^2	r squared (variance explained)
R^2_a	adjusted r squared (also reported as $R^2_{adjusted}$)
r_s	Spearman correlation coefficient
Q	Interquartile range
s	Number of points
SE	standard error
SD	Standard deviation
SRMR.....	standardized root mean square of residual
t	t statistic

TLI.....	Tucker-Lewis index
U	Mann-Whitney test statistic
VIF.....	Variance inflation factor (for regression models)
W	test statistic of Shapiro-Wilk test (normal distribution)
χ^2	test statistic of Mauchly's sphericity test
χ^2	chisquare
χ^2/df	ratio of chisquare and degrees of freedom
z	z-score
Z	test statistic of Wilcoxon test

1 Introduction

Research on vocabulary knowledge and its potential predictors has a long research tradition (e.g. Nation, 2013; Schmitt, 2010). To examine vocabulary knowledge more closely several vocabulary tests have been developed and are widely in use (e.g. Nation, 1990; Schmitt, Schmitt & Clapham, 2001). However, while most of these vocabulary tests focus on the knowledge of words which are highly frequent in a general context, few measurement instruments examine knowledge of words which are highly salient in an academic setting. Moreover, most vocabulary tests in use have not attempted to measure more than one word knowledge aspect, such as the link between a word form and its primary meaning or collocational knowledge. Consequently, only few vocabulary test batteries covering different word knowledge aspects of the same target words exist (e.g. González-Fernández & Schmitt, 2019; González-Fernández, 2021). Therefore, one aim of this study is to develop a vocabulary test battery, consisting of measurements of four word knowledge aspects of the same academic verbs (see chapter 7) on the basis of testing and assessment theory (see chapter 3.1).

This vocabulary test battery is consequently used to provide insights into English major students' knowledge of these academic verbs (see chapter 8) and how their knowledge develops over one academic year (see chapter 8.4). While prior studies investigating general and academic vocabulary knowledge development exist (e.g. Luxton, Fry & Coxhead, 2017; Townsend et al., 2016), these studies are mostly pseudo-longitudinal studies, while this study has adapted a longitudinal study design following the same participants for one academic year.

As especially the last decades have been very fruitful concerning insights into the nature of vocabulary knowledge, establishing several approaches to defining and reflecting vocabulary knowledge (e.g. Milton & Fitzpatrick, 2014; Nation, 2013) as outlined in chapter 2, this study is further striving to contribute to this research area. Building on prior studies, such as González-Fernández and Schmitt (2019), this study tries to come up with a model for vocabulary knowledge including the four word knowledge aspects researched using structural equation modelling. These are form-meaning recognition, form-meaning recall, collocation recognition and multiple meaning recognition knowledge, which are further described in chapter 2.4.2.1. Furthermore, it is attempted to show whether such a vocabulary model and the relationships between the word knowledge aspects researched are stable over the time of investigation. These findings can be found in section 8.5.

Zooming in on research on potential positive predictors of vocabulary knowledge reveals that most prior research has focused on the influence of intentional learning on vocabulary knowledge. However, since English has become a global language and lingua franca in the past decades (e.g. Seidlhofer, 2005), the language has managed to escape the walls of

classrooms and has successfully entered peoples' lives outside of an educational setting. This process and its impact on language proficiency has been investigated by the field of Extramural English, described in more detail in chapter 0. Working in this research field this study is striving to provide further insights into Extramural English practices of English major students in an Austrian setting, hoping to add to already existent knowledge about Extramural English behaviour at various educational levels. Furthermore, while prior Extramural English studies have been largely cross-sectional this research aims to add longitudinal information about Extramural English behaviour and potential trends related to biographical background variables. These findings are reported in chapter 9.1 and 9.7.

Trying to combine this research area with the data collected regarding vocabulary knowledge, this study examines whether Extramural English behaviour of English major students in Austria is linked to their academic vocabulary knowledge. While several other studies have already examined a potential link between Extramural English and knowledge of general vocabulary, only few studies (e.g. Busby, 2020; Lee, 2019a; Olsson & Sylvén, 2015) have investigated the relationship between Extramural English and academic vocabulary knowledge. Additional to this research gap this study further tries to fill the dearth of research regarding the influence of Extramural English on different individual word knowledge aspects. Consequently, this study provides information about the link between Extramural English activities and English major students' form-meaning recognition, form-meaning recall, collocation recognition and multiple meaning recognition knowledge of academic verbs, which are presented in section 10. Moreover, an interpretation of results and potential implications of findings also for an educational setting are offered in chapter 11.

2 Vocabulary knowledge

As González-Fernández and Schmitt (2019) outline in their study of vocabulary knowledge components, "vocabulary knowledge is a complex construct" (p. 481), which deserves thorough theoretical and practical investigation. This already gives away the purpose of this chapter: Read (2000, p. 16) states "before we start to consider how to test vocabulary [knowledge], it is necessary first to explore the nature of what we want to assess". Starting with trying to answer the simple sounding question "what is a word?", this chapter attempts to introduce the relevant theoretical concepts and terms linked to vocabulary knowledge underlying this research project. This includes the topicalization of multiword units, different types of words and finally what it actually means to "know" a word.

2.1 Definition of a word

While it appears to be a common assumption that assessment of vocabulary knowledge means testing the knowledge of words (Read, 2000, p. 17), some researchers already

disagree concerning the use of the term *word* (Miralpeix, 2020, p. 191). Hence, a synonymous use of the terms *lexical item* (e.g. Carter, 1987) or *unit of meaning* (Sinclair, 2004) can be found in the literature. Even more variation is visible when looking at the definition of the unit *word*, depending on the sub-discipline considered (Gardner, 2007).

From an orthographic point of view a word can be defined as “the black marks you are reading on this page and you know when one word ends and another one begins because there are spaces between words” (Daller, Milton & Treffers-Daller, 2007, p. 2; Miralpeix, 2020). While this definition might dispose a specific everyday useability, three main problems can be identified, namely variability of orthography, treatment of compounds and the fact that also illiterate speakers are able to distinguish between units of language (Plag, 2003). Hence, when using an orthographic definition of a word it has to be clarified in advance which orthographic rules are adhered to and whether compounds, such as *high-rise apartment*, are considered as one or two words. Similarly a word can be described as a unit of speech carrying one main stress and consisting of several individual sounds (Plag, 2003, p. 6). Nevertheless, this accounts only for content words, which are not reduced to clitics, since function words and clitics carry no stress. Moreover, when looking at a word from a morphological perspective it can be defined as “[...] an indivisible unit into which no intervening material may be inserted”, which is referred to as *integrity criterion* (Plag, 2003, p. 6). While this holds true for most words such as *girl-s* or *un-happy*, compounds as *son-in-law* pose an exception, since the plural marker is inserted in the word in *sons- in- law*. Semantically a word could be claimed to be “one or more words that together make up a unit of meaning” (Sinclair, 2004, p. 281), while syntactically they are considered to be “syntactic atoms” (Plag, 2003, p. 7), classified as *parts of speech*, which are subjected to syntactic rules. While all fields offer new perspectives on the concept of a word, these morphological, phonological and syntactic properties have been accepted as universal properties of a word:

- words are entities having a part of speech specification
- words are syntactic atoms
- words (usually) have one main stress
- words (usually) are indivisible units (no intervening material possible)

(Plag, 2003, p. 8)

In the field of vocabulary assessment, which aims at establishing a counting unit to measure the number of words known in a language, further differentiations into tokens, types, lemmas and word families have been made.

Tokens “refer to the total number of words in a corpus” (Brysbaert et al., 2016, p. 1), which corresponds to counting each word in a text (Nation, 2013, p. 9). For instance, the sentence *I think I need to take my dog for a walk*, consists of eleven tokens but only ten types, since when using the counting unit types words are not counted again when they occur more

than once (Nation, 2013, p. 9). Hence, the word *I* is only counted once in the example. As Read (2000, p. 84) states both tokens and types are potential counting units for estimates of vocabulary knowledge. However, vocabulary knowledge has also been estimated using larger units, namely lemmas and word families.

Lemmas can be defined as “[including] both the root word and any morphological derivatives in the same grammatical class, described as inflectional morphemes” (Duff & Brydon, 2020, p. 457). For example the lexeme of the root word *dance* (verb) would include the inflectional morphemes *dances*, *danced* and *dancing*. Moreover, the noun *dance* would be excluded from this lemma because it belongs to another grammatical class. It is assumed that if a learner knows the root word of a lemma, he or she can be assumed to know the inflectional morphemes encompassed in this lemma as well. Hence, it is argued that counting each of the words subsumed under a lemma as individual word would result in an inflated estimate of vocabulary knowledge (Goulden, Nation & Read, 1990, p. 343). Consequently, some vocabulary tests calculate estimates of lemma knowledge by testing the root words of lemmas, assuming that the morphological derivatives in the same grammatical class are known as well (Miralpeix, 2020).

Sometimes an even larger counting unit than the lemma has been used to estimate vocabulary knowledge, namely the unit of word families. A word family “is a unit that includes both the base form of the word and its inflections, as well as its main derivations” (Vilkaitė-Lozdienė & Schmitt, 2020, p. 83). For instance, the word family of the base form *work* would consist of the inflections *works* and *worked* as well as derivations such as *worker*. Hence, words grouped in the same word family are not restricted anymore by derivative forms of the root word belonging to the same grammatical class as it was the case for the lemma. Nevertheless, similar to the counting unit lemma it is assumed that learners do not have to invest the same amount of learning effort into the acquisition of inflected and derived forms of a word as soon as they have acquired the base form (Read, 2000b, p. 85; Vilkaitė-Lozdienė & Schmitt, 2020, p. 84). However, crediting a learner the knowledge of all word family members rests on the assumption that a learner acquires the base form of a word family first and is able to deduce the meaning of related forms from it (Gardner, 2007, p. 248; Nation, 2021, p. 969). Yet, there is no evidence that learners might not encounter any of the derived forms first when being exposed to authentic input and whether they would equally be able to derive the meaning of the base form of the word family from an initially learned derived word form (Biemiller & Slonim, 2001). Furthermore, research has shown that “age, English literacy skills, and extent of morphological training [...] affect the ability of English language users to recognize and utilize morphological relationships between words” (Carlisle, 2000; Freyd & Baron, 1982; Gardner, 2007, p. 258; Mahony, Singson & Mann, 2000; McBride-Chang et al., 2005; Tyler & Nagy 1990). Therefore, learners with different preconditions also differ regarding the extent to which they can be credited the knowledge of related word forms when knowing a base form. Hence, Gardner

(2007, pp. 258-259) recommends using different counting units to measure the vocabulary knowledge of different groups:

- a) *Base forms + regular inflections*: younger children; low general English proficiency; low English literacy skills; no specific morphological training
- b) *Base forms + regular inflections + irregular inflections + derivational prefixes*: older children and adolescents; intermediate general English proficiency; intermediate English literacy skills; some morphological training
- c) *Base forms + regular inflections + irregular inflections + derivational prefixes + derivational suffixes (regular) + derivational suffixes (irregular)*: adults; high general English proficiency; high English literacy skills; extensive morphological training or experience

According to these guidelines, vocabulary tests, which base their estimates of vocabulary knowledge on the counting unit of word families or lemmas are only suitable for intermediate to advanced learners, while vocabulary tests basing their count on individual types can be utilized for a broader audience. Since, it is not clear how much morphological training or experience the participants of this study have, it was decided that the vocabulary tests used for this research base their estimates of vocabulary knowledge only on the knowledge of the types tested and their regular inflections regarding recall knowledge and do not assume knowledge of derived forms beyond this level.

2.2 Multiword units

While research on individual words or lexical items as defined above has been flourishing for decades, the fundamental role of the knowledge of multiword units has long been neglected (Wray, 2002, p. 4), accompanied by a lack of research on multiword units (Read, 2000a, p. 23). However, more recently “research is increasingly showing that [...] vocabulary is often made up of formulaic multi-word sequences” (Li & Schmitt, 2009, p. 85). Moon (2011, p. 43) broadly defines a multiword unit as “a vocabulary item which consists of a sequence of two or more words. This sequence of words semantically and/or syntactically forms a meaningful and inseparable unit”. While this general definition allows a first impression of the concept, Nation (2013) emphasizes that several answers to the question *what is a multiword unit* exist, which are dependent on the perspective from which you are investigating the phenomenon.

Similar to Moon’s (2011) approach he states that multiword units can be “all the combinations of a particular word or type of word and its accompanying words, whether they are highly frequent, strongly associated or not” when being solely defined by their form (Nation, 2013, p. 479). However, very frequently multiword units are characterized as “a group of words that commonly occur together” when addressing the issue from a frequency-based perspective (Nation, 2013, p. 479). Using this definition means that

multiword units are combinations of words, such as *take a chance*, which occur very frequently with each other in a language. Alternatively, the term multiword unit has often been related to the concept of compositionality, therefore being “a group of words where the meaning of the phrase is not obvious from the meaning of the parts”, such as *kick the bucket* (Nation, 2013, p. 479). Finally, when coming from the field of psycholinguistics multiword units are said to be stored and processed as single lexical items (Conklin & Schmitt, 2008; Nation, 2013).

When considering these different approaches to the concept of multiword units it does not come as a surprise that the field is characterized by inconsistency and non-transparency regarding the use of terminology (Biber, 2006; Schmitt & Carter, 2004; Wray, 2002). Wray (2002, p. 9) lists around 50 terms referring to the concept of formulaic language and multiword units, including *lexical bundles, collocations, phrasal units, chunks, formulas or multiword items*. One potential reason for this flood of terminology might be the multi-faceted nature of the concept of multiword units, consisting of a “surprising range and scope of formulaic language” (Wood, 2019 p. 30). While acknowledging that up to date attempts towards a categorization of different types of multiword units still suffer of imprecision, resulting in overlapping categories with fuzzy boundaries, Wood (2019, p. 30) attempts to list different types of multiword units in an organized manner. He distinguishes types of multiword units which are “characterized mainly by their structural/semantic/syntactic properties” from types of multiword units which are predominantly characterized by their “pragmatic utility” and types of multiword units being dependent on “their distribution in particular corpora” (Wood, 2019, p. 31).

In Table 2.1 the different types of multiword units are presented together with a very short definition and an example. More attention will be paid to the concept of *collocation* in section 2.2.1, since it is one of the concepts investigated in the scope of this research study.

Type of multiword unit	Definition	Examples
<i>Sequences Distinguished by Structural, Semantic, or Syntactic Properties</i>		
Collocation	“the occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991, p. 170).	<i>make a decision, give a speech, hard work, bated breath</i>
Idioms	“[...] multi-word items which are not the sum of their parts: they have holistic meanings which cannot be retrieved from the individual meanings of the component words” (Moon, 2011, p. 46).	<i>spill the beans, kick the bucket, have an axe to the grind, rain cats and dogs</i>
Metaphors	“A metaphor is a sequence based on an unconventional reference in which words are used to	<i>time is a healer, life is like a box of chocolates</i>

	describe something ordinarily far from its normal scope of denotation, producing a discord between a literal interpretation and a metaphoric interpretation” (Wood, 2019, p. 33).	
Proverbs	“Proverbs are generally sentence-length sequences which display an opaque relationship between literal and figurative meanings” (Wood, 2019, p. 33)	<i>a stich in time saves nine, early to bed and early to rise, like death and taxes</i>
Compounds	“A compound is a sequence of two words, the second of which usually functions as the head or core of the compound” (Wood, 2019, p. 34).	<i>desk computer, hardcore, brother-in-law, grocery store</i>
Phrasal verbs	“Phrasal verbs are distinguished largely by their distinctive structural makeup, lexical verbs combined with a preposition, particle, or both, with often nonliteral meanings, or both literal and figurative interpretations, like idioms” (Wood, 2019, p. 34).	<i>pick up, bump into, show off, carve in, looks up</i>
<i>Sequences distinguished by pragmatic utility</i>		
Lexical phrases	“Lexical phrases are a pragmatically specialized subset of formulaic sequences [...]. The phrases fall into two structural categories: strings of specific lexical items [...] and generalized frames. There are four broad categories of lexical phrases: polywords [...], institutionalized expressions [...], phrasal constraints [...] and sentence builders” (Wood, 2019, p. 35).	<i>for the most part, nice to meet you, a...ago, I think that...</i>
Pragmatic formulas	“In pragmatics, formula is used to refer to formulaic sequences employed for specific pragmatic purposes. [...] These include greetings [...] and turn-taking” (Wood, 2019, p. 35).	<i>how are you, what’s going on, how are things, let me add, I also think</i>
<i>Sequences distinguished by their distribution in corpora</i>		
Lexical bundles	Lexical bundles can be briefly defined as “combinations of three or more words which are identified in a corpus of natural language by means of corpus analysis software programs” (Wood, 2019, p. 35).	<i>the size of the ..., one of the things, ...are likely to be..., as well as...</i>
Concgrams	“A concgram is a combination of two or more words, but is distinctive in that it is a noncontinuous sequence, with the constituent words separated by others” (Wood, 2019, p.36).	<i>And...then</i>

Table 2.1: Types of multiword units

While considerable disagreement can be found concerning the use and definition of terminology in the field, there is more consensus on the importance of multiword units and their benefits. Many researchers (Conklin & Schmitt, 2008; Conklin & Schmitt, 2012;

González-Fernández & Schmitt, 2015; Hong & Hua, 2018; Schmitt & Carter, 2004) agree on the fundamental role of multiword units for the attempt to achieve native-like mastery in a language. More specifically multiword units appear to aid fluency in spoken (Conklin & Schmitt, 2008; Conklin & Schmitt, 2012; González-Fernández & Schmitt, 2015) and written form (Schmitt & Carter, 2004), which might be attributed to the fact that multiword units are stored and processed as single choices in the mental lexicon, allowing for quicker processing and retrieval (Arnon & Priva, 2013; Conklin & Schmitt, 2008; Conklin & Schmitt, 2012; Ellis et al., 2008; Jiang & Nekrasova, 2007; Nation, 2013; Schmitt & Carter, 2004; Siyanova & Schmitt, 2007). Hence, “the language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair, 1991, p. 110). This phenomenon was described by Sinclair (1987) as the so called *idiom principle*, which states that “the choice of one word to begin to express an idea has a strongly limiting effect on what the following word choices can be” (Nation, 2013, p. 483). Therefore, one can say that lexical choices are more limited as one would maybe expect when applying the *open-choice principle*, which mainly states that “sentences are produced creatively on the basis of an underlying system of rules” (Read, 2000a, p. 21). Sinclair (2004, p. 280) goes even so far to claim “that most lexical meaning is associated with word patterns rather than with individual words”, which he subsumes under the so called maximal approach (Gardner, 2007, p. 255). Wray (1992) proposes that the principles a language user adheres to will depend largely on the situation and communicative needs. Hence, ‘real’ language use consists of a balancing act between these principles.

Taking a more pedagogical perspective on the storage of multiple word items as single units reveals that this processing advantage is only available for advanced L2 learners to a similar extent as for native speakers (Conklin & Schmitt, 2012; González-Fernández & Schmitt, 2015; Paquot & Granger, 2012). Mid- or low-proficiency students appear to deconstruct multiword units word by word (Conklin & Schmitt, 2012, p. 45), forfeiting the benefits of the *idiom principle*. Consequently, research has found that L2 learners use fewer multiple word units than native speakers (Granger, 1998; Howarth, 1998a; Laufer & Waldman, 2011), with some learners even consciously avoiding (Gardner, 2007; Liao & Fukuya, 2004) or overusing familiar, often atypical, multiword items (González-Fernández & Schmitt, 2015, p. 97; Laufer & Waldman, 2011). For instance, Aijmer (2009, p. 153) in her study of Swedish EFL learners has found that her participants overused the phrase *I don’t know*, applying it mostly as speech management signal, while native speakers use it predominantly to avoid posing a direct question. Nesselhauf (2005, p. 69) has referred to this phenomenon of overuse as “collocational teddy bears”, which learners feel safe using. However, the lack of multiword units in L2 learner language cannot solely be attributed to conscious avoidance or overuse of familiar items but it needs to be acknowledged that multiword knowledge tends to “[lag] behind their general language and vocabulary knowledge” (González-Fernández & Schmitt,

2015, p. 94; Irujo, 1993). Consequently, it needs to be examined how the knowledge of multiword units within an L2 context can be increased and fostered through implicit and explicit approaches, which will be thematized in the respective sections 0 on Extramural English and vocabulary learning and in results chapter 10.1.

While the degree to which language is composed out of free decisions of individuals and pre-existent chunks generated over time through habits of language users is still under investigation, it has been found that large proportions of spoken and written text consist of multiword units (Altenberg, 1998; Erman & Waren, 2000; Gardner, 2007; Paltridge, Starfield & Coxhead, 2017). However, estimates vary from 80% of the London Lund Corpus of Spoken English (Altenberg, 1998) to 52-58% in Erman & Warens' study of English written texts in 2000. Similar to estimates of single lexical units, these differences may at least partly result from divergent counting units chosen (Paltridge, Starfield & Coxhead, 2017). However, it is without doubt that L2 learners would benefit from the acquisition of multiword units, since they make up a large proportion of the English language, and learning them would enlarge their fluency in both spoken and written form. Apart from these advantages multiword units have additionally shown a large pragmatic potential (Conklin & Schmitt, 2008; Hong & Hua, 2018; Nation, 2013).

She would go and smile and be nice and say 'So kind of you. I'm so pleased. One is so glad to know people like one's books'. All the stale old things. Rather as you put a hand into a box and took out some useful words already strung together like a necklace of beads. (Agatha Christie, *Elephants Can Remember*)

As visible in Agatha Christie's quote above, multiword units are "often used to accomplish recurrent communication needs", which make use of conventionalized language (Conklin & Schmitt, 2008, p. 73). For example multiword units are used to accomplish the speech acts of expressing sympathy (*I'm (very) sorry to hear about*), answering a request (*I'd be happy/glad to*), making requests, giving directions or maintaining social interaction (Schmitt & Carter, 2004, p. 9). Schmitt and Carter (2004, p. 10) emphasize that language users tend to apply multiword units when engaging in everyday interactions, which primarily serve the purpose of keeping a conversation going and are not focused on particular content. For instance, comments about the weather (*Nice weather today; Cold isn't it*), backchanneling (*Did you really?; How interesting*) and communicating understanding (*Oh, I see what you mean; OK, I've got it*) are highly formulized (Schmitt & Carter, 2004, p. 10). However, multiword units additionally play an important role concerning structuring of discourse (Conklin & Schmitt, 2008; Schmitt & Carter, 2004; Wray, 2002). For example Drew and Holt (1998) have found that they "occur at places of topic-transition and as summaries of gist" (Conklin & Schmitt, 2008, p. 74). More concretely multiword units tend to be used when opposing perspectives are voiced (*on the one hand...on the other hand*), when signposts are made to other sections in the discourse (*as I was saying, speaking of which*) or when repeating information in a paraphrased manner (*to put it another way*) or to summarize

information (*as discussed in, it has been suggested*) (Paquot & Granger, 2012, p. 134; Schmitt & Carter, 2004, p. 10). Moreover, multiword units are an efficient way to avoid misconception in a specific technical field. As Schmitt and Carter (2004, p. 10) exemplify nicely blood pressure is always reported in the following manner *blood pressure is 140 over 60*, with the higher value named first, which prevents misunderstandings. Therefore, multiword units can be seen to serve several important pragmatic functions, highlighting their importance for a language learner even more.

On a similar strand several researchers have argued that multiword units can be said to have semantic prosody, similar to single lexical units (Conklin & Schmitt, 2008; Nation, 2013; Schmitt & Carter, 2004; Wray, 2002). This means that multiword units have “a type of register marking” (Conklin & Schmitt, 2008, p. 74), which allows a language user to show his or her stance, evaluation or attitude. Schmitt and Carter (2004, p. 74) exemplify their point by stating that phrases like *I don’t know if...*, or *I want you to ...* allow language users to show their stance and wishes. Nation (2013, p. 481) summarizes this phenomenon nicely by concluding “that words need to be learned in their typical multiword units, largely because their most common senses are not what we would intuitively expect ([e.g.] the most common sense of *see* is understand and not see visually) and their frequent multiword units reveal this”.

Considering all benefits of multiword units named above, it becomes obvious that the knowledge of multiword units needs to be addressed in current research and that assessment of vocabulary knowledge should not be restricted to estimates of individual lexical items anymore. Therefore, it was decided to include the concept of receptive collocational knowledge in this research project, which is why the term collocation will be topicalized in more detail in the following.

2.2.1 Collocations

While the term collocation goes back to the 1600s, it was not until 1931 that Palmer (p. 4) used the term to denote “units of words that are more than single words”. However, similar to the overarching concept of multiword units, the field of collocation research is characterized by the lack of a consistent definition (Coxhead & Byrd, 2012; Gyllstad, 2007). Nevertheless, two prominent positions can be located, namely the phraseological and the frequency based approach (Barfield, 2006; Bonk, 1996; Ebrahimi, 2018; Gyllstad, 2007).

The phraseological approach to collocation research is mainly dominated by the work of Cowie (1988, 1991, 1994, 1998a, 1998b, 1998c), Howarth (1996, 1998a, 1998b), Melčuk (1998) and Benson, Benson and Ilson (1997), while building on work of Palmer (1931, 1933). As Gyllstad (2007, p. 15) summarizes, the phraseological tradition is united by their view that collocations “are some kind of word combination, displaying various degrees of fixedness”, coming up with various theoretical approaches and models for a better understanding of the phenomenon. Figure 2.1 below depicts how a model uniting the views

of Cowie (1988, 1991, 1994, 1998a, 1998b & 1998c), Howarth (1996, 1998a, 1998b) and Benson, Benson and Ilson (1997) could look like.

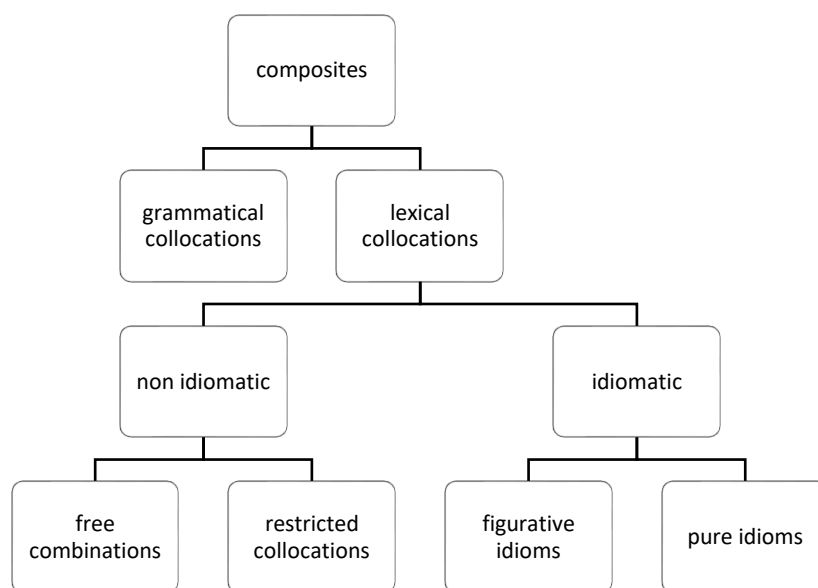


Figure 2.1: Model of phraseological approach to collocation combining Cowie (1994, 1998a, 1998b), Howarth (1998a, 1998b) and Benson et al. (1997)

Initially, Cowie (1998a, 1998b) distinguished between so *called formulae and composites*. The former being “units of sentence length” and the latter being “units from below the sentence level” (Gyllstad, 2007, p. 11), which encompasses the concept of collocations. Building on the theoretical groundwork of Cowie (1998a; 1998b), Howarth (1996, 1998a, 1998b) further distinguishes between grammatical and lexical composites, which are determined by the word class of the constituent word (Gyllstad, 2007), with grammatical composites being combinations including function words and lexical composites comprising only content words. Although the terms grammatical and lexical composites, including the term composite, imply that not only collocations, but also other composite types could be affected by this classification, it can be assumed that Howarth’s (1998a, 1998b) categorization is comparable to Benson, Benson and Illsons’ (1997) differentiation between grammatical and lexical collocations. Additionally, he further distinguishes between non idiomatic lexical composites and idiomatic lexical composites for lexical collocations/composites. This is nicely supplemented by Cowie’s (1994) first attempt to introduce a scale of fixedness or variability, which resulted in the categories of free combinations on the one end, followed by restricted collocations, figurative idioms and pure idioms on the other end of the continuum. However, only free combinations and restricted collocations are relevant for the concept of collocational knowledge and have stayed in use over time (e.g. Paquot & Granger, 2012). While free combinations are completely variable and transparent word combinations as *eat one’s food* (Cowie, 1994), restricted collocations can be described as “lexically constrained combinations that allow for limited substitution within a particular grammatical construction” (Paquot & Granger, 2012, p. 136). However, it

has to be acknowledged that the distinction of non idiomatic and idiomatic should be regarded as continuum from free combinations, over restricted collocations, figurative idioms to pure idioms at the extreme end of the scale and not as strict division into separate categories.

While the phraseological approach follows a rather theory-oriented deductive concept the frequency based approach sees the concept of collocations as “units consisting of co-occurring words at a certain distance from each other, and a distinction is often made between frequently and infrequently co-occurring words” (Nesselhauf, 2005, p. 11). As for the phraseological tradition the frequency based approach was coined by several key figures, namely Firth (1951, 1957, 1968), Halliday (1961, 1966) and Sinclair (1966).

While Firth (1951, 1957, 1968) laid the groundwork by making the concept of collocation more widely known, Halliday (1961, 1966) attempts to narrow the frequency based concept of collocation further by stating that a collocation can be defined as “the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c” (Halliday, 1961, p. 276). When looking at this definition in more detail one can deduce that he leaves the number of words involved in a collocation open by stating that it can be n words. This number of words encompassed in a collocation is called span. Moreover, Halliday (1966) introduces the term node for the target word which is under investigation in a collocation. The words surrounding the node are called collocates.

In addition, it can be said that Halliday (1961) was the first to thematize the statistical nature of collocates by stating that they are quantifiable. He sees collocations as the probability that one word occurs together with (an)other word(s) in a certain distance of each other. Building on this mathematical approach Sinclair (1966) is the first to introduce a formula to calculate collocational strength, which has been supplemented by more statistical measures to calculate collocational strength, such as the Mutual information (MI) score, the z-score, the t-score, raw frequency or loglikelihood measure over the time. Out of these the z-score and the MI score have been used in this study and will therefore be described in more detail below.

z-score

Similar to other statistical measures the z-score is used to indicate how likely it is for a node word and a collocate word to occur together in a given corpus of texts (Gyllstad, 2007). If we want two words to be very likely to co-occur, the probability of them occurring together by chance should be rather high. This is indicated by the so called p-value, which denotes the percentage by which a given result is due to chance. While there are different cut off levels for the p-value, a boundary of 0.05 is common in social sciences (Gyllstad, 2007). This means that if a p-value is lower than 0.05, the chance of the result being by chance and not purpose is even smaller than five percent (Gyllstad, 2007).

Building on the p-score the z-score indicates whether the probability of one word occurring with another word in a specific distance is larger than chance. This is done with the following formula by Oakes (1998, p. 163):

Z = number of words in a given corpus

A = how often the node occurs in the corpus (F_n)

B = how often the collocate A occurs in the corpus (F_c)

O = number of co- occurrences of node and collocate

S = span size of the collocation

First the probability (p) of the collocate to occur where the node is not is calculated:

$$p = \frac{F_c}{(Z - F_n)}$$

Then the expected number of co-occurrences of the node and the collocate (E) is calculated:

$$E = p * F_n * S$$

These values are then inserted into the formula for the z-score:

$$Z = \frac{(O - E)}{E * (1 - p)}$$

MI score

In contrast to the z-score the MI-score does not identify collocations which are very frequent in a given language corpus but locates collocations which have a particularly strong relationship to each other (González-Fernández & Schmitt, 2015). This is attempted by calculating “how likely a given set of words are to occur together in a set sequence in comparison to chance” (Wood, 2019, p. 38). Hence, the MI-score indicates the “strength of association” instead of a mere frequency count (Ackermann & Cheng, 2013, p. 237). While there is no consistent boundary when a collocation might be considered as significant when using the MI score, several studies (e.g. Church & Hanks, 1990; Ebrahimi, 2018; Hunston, 2002) have used a cut off level of an MI score larger than 3.0 (Wood, 2019). This approach has also been adopted for this study, using the MI score in combination with the z-score to arrive at an objective measure of probability of co-occurrence.

Placing this research study according to the above mentioned theoretical approaches will most likely result in a more frequency-based approach to the concept of collocations, since statistical measures have been consulted to identify meaningful collocations to include in the vocabulary tests designed. However, the choice of target collocations was not only guided by frequency-based considerations, as form- and meaning-based criteria were taken into account as well (see Table 2.2).

Form-based factors		
Adjacency Words in the multiword unit can occur right next to each other.	vs.	Discontinuity Words in the multiword unit can be separated by a word or words not in the multiword unit.
Grammatically fixed A multiword unit can be counted as a fixed form which is unchanging.	vs.	Grammatically variable A multiword unit can occur together in a variety of grammatical and affixed forms.
Grammatically structured A multiword unit can be a complete grammatical unit, such as a sentence, a sentence subject, a predicate, an adverbial group etc.	vs.	Grammatically incomplete A multiword unit can be grammatically incomplete.
Lexically variable Counting multiword units can include some substitutable words of related or similar meaning.	vs.	Lexically invariable Counting multiword units can contain items that cannot be replaced by others.
Number of components A multiword unit must contain at least two words. Some counting is done with a limit on the number of words in the multiword units, some counting has no limit.		
Meaning-based factors		
Compositionality The meaning of a multiword unit can be the sum of the meanings of its collocates.	vs.	Non-compositionality The meaning of a multiword unit can be more than the sum of the meanings of its collocates.

Table 2.2: Form-based and meaning-based factors affecting multiword unit selection (Nation, 2013, p. 488)

Following Nation's (2013, p. 488) list of form-based criteria, which might influence what is counted as a multiword unit and therefore the selection of target items, it was decided that multiword units consisting of only two components should be included. Moreover, all multiword units chosen were adjacent; hence occurring directly next to each other to avoid uncertainty regarding the affiliation of potential collocates to the node word. Regarding grammatical fixedness or variability all items were grammatically variable, consisting of verb+noun or noun+verb collocations, which can be found in different derivational forms. For instance, the collocation *approach retirement* also works with different affixes attached, such as *approaches retirement*, *approached retirement*, *approaching retirement*. However, collocations occurred most often in the base form, being the uninflected verb+noun collocation. Concerning the notion of grammatical structuredness or incompleteness, grammatically complete items being a predicate and an object were chosen. Finally, the last form-based factor considered, namely whether collocates are lexically variable or invariable, already leads over to meaning-based criteria for choice of multiword units. All multiword units in this study were compositional regarding their meaning, which means that the meaning of the multiword unit is deducible from the individual parts. Moreover, multiword units were lexically variable in the sense that the verbs can be found in combination with other nouns. However, no combinations with the same salient meaning were possible.

Apart from form-based and meaning-based criteria influencing the type of multiword unit chosen, another restricting criterion, namely a particular type of words, was applied, which will be addressed in the next chapter.

2.3 Types of words

Acquiring the vocabulary of a new language is definitely a crucial part of mastering this language (Brezina & Gablasova, 2015; Ishihara, Okada & Matsui, 1999; Read, 2000; Vilkaitė-Lozdienė & Schmitt, 2020). Nevertheless, it has to be considered that language learners pursue different purposes why they want to learn a new language. While some might start learning a language at school out of general interest, others might strive to add a more academic touch to their English writing or might want to become more proficient in a particular target area, such as medicine or law. Hence, the choice of vocabulary items relevant for the learner should be related to the initial motivation or the intrinsic aim of the particular person (Vilkaitė-Lozdienė & Schmitt, 2020). For this purpose the concept of word frequency has proven to be a useful tool, which will be elaborated in the following.

2.3.1 High-frequency vocabulary

As research has shown, some words of the English language are used more frequently in written and/or spoken discourse than others (Nation, 2013; Nation & Waring, 2011; Schmitt & Schmitt, 2014). For instance, Nation (2006, p. 79) shows that 78-81% of all tokens in the British National Corpus (BNC) can be attributed to the most frequent 1000 word families of

the English language. Consequently, it can be argued that “frequent words are more important [for language learners], as they are encountered more often than less frequent words” (Vilkaitė-Lozdienė & Schmitt, 2020, p. 82). These frequent words have been named high-frequency words by Nation (2013), who defines them as the most frequent 3000 word families of the English language comprising both content and function words.

The investigation of high-frequency words, being most useful for learners at a beginner stage, already dates back to the early 20th century when first attempts were made to structure the vocabulary of the English language, which is known as the Vocabulary control movement (Vilkaitė-Lozdienė & Schmitt, 2020). Publications, such as The Teachers’ Word Book of 30 000 words (Thorndike & Lorge, 1944) or the American Heritage Word Frequency Book (Carroll, Davies & Richman, 1971), were amongst the first compilations of high-frequency words in form of a list of word families (Nation & Waring, 2011; Vilkaitė-Lozdienė & Schmitt, 2020).

One of the most influential lists of high-frequency words is West’s (1953) General Service List (GSL), presenting around 2000 of the most frequent word families of the English language (Nation, 2013; Nation & Waring, 2011; Read, 2000; Schmitt & Schmitt, 2014; Vilkaitė-Lozdienė & Schmitt, 2020). While the GSL has been frequently utilized for research in the area of vocabulary development and vocabulary testing in the past, several researchers have criticized recently that the GSL is outdated (Gardner & Davies, 2014; Nation & Waring, 2011; Read, 2000; Vilkaitė-Lozdienė & Schmitt, 2020). This claim is supported by the fact that the GSL includes several words, which are not in use to the same extent anymore [e.g. *gay=happy, shilling, cart, footman, milkmaid*], while excluding words, which are frequent in our present day language use [e.g. *television, mobile phone, internet*] (Brezina & Gablasova, 2015). Moreover, West (1953) has used more subjective selection criteria apart from the objective criteria of word frequency to compile the GSL (Brezina & Gablasova, 2015; Read, 2000), such as including less frequent words in the list, if they show similarities to another word form, since he thought that they would have a particularly low learning burden. In addition, early approaches at high-frequency lists, such as the GSL (West, 1953), are still based on relatively small text corpora, since the compilation of the lists was still done by hand (Vilkaitė-Lozdienė & Schmitt, 2020). It is only with the emergence of computer based corpora that large numbers of text could be analyzed automatically (Vilkaitė-Lozdienė & Schmitt, 2020).

Considering these points Brezina & Gablasova (2015, p. 2) have developed the “New General Service List”, which is based on four different corpora of varying size and age, namely the LOB, BNC, BE06 and the EnTenTen12, with a total size of over twelve billion tokens. The New General Service List (Brezina & Gablasova, 2015, p. 1) includes the “stable vocabulary core of 2122 items (70.7% among the four corpora [...]) and covers between 80.1 and 81.7 per cent of the text in the source corpora”. However, research has indicated that

the GSL (West, 1953) has a large overlap with comparable more recent lists (Gilner & Morales, 2008; Nation, 2004; Nation & Hwang, 1995), implying that there appears to be a stable core of high-frequency words, which remains largely unchanged over time (Brezina & Gablasova, 2015).

2.3.2 Low-frequency vocabulary

Having defined the concept of high-frequency vocabulary, differing approaches to less frequent words have emerged. While initially all word families apart from the most frequent 3000 were categorized as low-frequency vocabulary, Schmitt and Schmitt (2014) have successfully challenged this approach by arguing that the words subsumed under this large concept are of varying utility for a language learner, which implies that further differentiation is needed. Recent research (Nation, 2013; Schmitt & Schmitt, 2014; Vilkaitė-Lozdienė & Schmitt, 2020) mostly adheres to the following model, which distinguishes between high-, mid- and low-frequency words (see Figure 2.2)

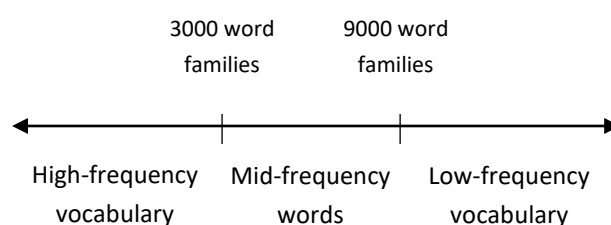


Figure 2.2: Types of vocabulary

To distinguish between mid- and low-frequency words Schmitt and Schmitt (2014) advocate a usage based approach due to the fact that both types of words do not considerably contribute to the percentage of words known in a text. Hence, Nation's (2006) corpus study was considered as reference, since it indicates that knowledge of 8000-9000 word families is needed to understand 98% of the running words in a variety of authentic texts. Davies (2008) supports this estimate by revealing that 95.5% of the Corpus of Contemporary American English (COCA) are covered by these 9000 most frequent word families. Therefore, the boundary of low-frequency words can be set "beyond the most frequent 9000 word families of English" (Nation, 2013, p. 28).

Within this group of low frequency words, items rarely used by nearly all language users, such as very old-fashioned, overly formal or vulgar terms [e.g. *eponymous*, *gibbous*] are clustered with proper nouns [e.g. *New York*, *Maria*] and words being strongly linked to a specific field [e.g. *scalpel*, *catheter*]. As Nation (2013, p. 29) highlights "[one] person's technical vocabulary is another person's low frequency vocabulary". Hence, low-frequency words, such as *scalpel* can be frequently used by members of a particular field or profession, such as surgeons, but are very infrequently used by the average language user. These terms are referred to as technical words and will be dealt with in more detail in section 2.3.4.

2.3.3 Mid-frequency vocabulary

Mid-frequency vocabulary, a term coined by Schmitt and Schmitt (2014), can be placed between high- and low-frequency words, comprising words, which belong to the 3000-9000 most frequent word families of the English language (Nation, 2013; Vilkaitė-Lozdienė & Schmitt, 2020). As research shows these mid-frequency words are needed in combination with high-frequency words to use English for authentic purposes without support (Nation, 2013; Schmitt & Schmitt, 2014). For instance, Webb and Rodgers (2009) found that it took 8000 word families for a coverage of 98% of all tokens in a news broadcast, 5000 word families to reach the same level of understanding for horror, drama and crime films and even 9000-10 000 word families for genres related to war and animation. Moreover, Laufer and Ravenhorst-Kaloski (2010) have concluded that 6000-8000 word families are needed for Israeli university students to master reading texts. Nation (2013) adds that 9000 word families are needed for proper understanding of English novels and 8000 for newspapers (98% coverage). Hence, mid-frequency vocabulary is central for the advanced mastery of the English language and should not only be seen as domain-dependent words tied solely to a person's interest (Vilkaitė-Lozdienė & Schmitt, 2020).

2.3.4 Technical vocabulary

Technical vocabulary (e.g. Chung & Nation, 2004; Nation & Coxhead, 2012; Nation, 2013), being "subject specific" vocabulary (DeChazal, 2014, p. 92; Paquot, 2010, p. 13), has been referred to under several different terms, such as "discipline specific vocabulary" (Woodward-Kron, 2008), "domain-specific glossaries" (Pernán-Pasqual, 2015), "scientific/technical terms" (Yang, 1986), "specialist terms" (Woodward-Kron, 2008), "specialized vocabulary" (Robinson, 1980) or "terminological" words (Becka, 1972; Kit & Liu, 2008; Peruzzo, 2014) or "units" (Cabré, 1999). This diversity of terminology is partially reflected in the definition of technical vocabulary, following two main approaches.

While the broader view on technical vocabulary admits that "technical words may occur in any subject" (Liu & Lei, 2020, p. 112), the narrow view confines technical vocabulary to the so-called hard sciences, such as engineering and medicine, and attributes specialised vocabulary to art, humanities and social sciences (e.g. Ardasheva & Tretter, 2017; Brieger & Pohl, 2002). However, this narrow view on technical vocabulary has become less common in recent times due to fuzzy disciplinary boundaries related to interdisciplinary work and increased incorporation of technology in humanity, arts and social sciences (Liu & Lei, 2020). Consequently, it is proposed that a broader view on technical vocabulary should be adopted. However, reviewing recent attempts to investigate the technical vocabulary of specific fields show that despite the inclusion of humanities, art and social sciences (e.g. Coxhead, Demecheleer & McLaughlin, 2016; Woodward-Kron, 2008) in the definition of technical terms mainly classical hard sciences have been researched. For instance, Perinan-Pascual (2015), Hsu (2013), Chen and Ge (2007), Wang, Liang and Ge (2008), and Chung and

Nation (2003) have examined the technical words of different medical fields, Yang (1986), Mudraya (2006) and Ward (2009) have focused on engineering and Ha and Hyland (2017) on finance. Only few studies, such as Vongpumivitch, Huang and Chang (2009), and Khani and Tazik (2013) have focused on the specific terminology of a humanistic field, namely Applied Linguistics. Therefore, a research gap regarding the technical vocabulary of humanity, art and social sciences can be located, which is even more pressing considering the crucial nature of technical terms for the language learner.

Within the technical vocabulary of a field, being “words and phrases that are used and known mainly in a specific [...] subject area” (Liu & Lei, 2020, p. 111), different perceptions of what should be counted as technical words exist. As research has shown (Lei & Liu, 2016; Nation, 2013; Nation & Kyongho, 1995) a large majority of technical words do not consist of low-frequency words with restricted use in this subject area but are also composed of high-frequency and academic words with and without a specialized, polysemous nature. Nation (2013) extrapolates his findings of an academic textbook analysis and reveals that 45% of the technical words found in a specialized text are high-frequency words, 30% are academic words and only 25% are technical words coming from the category of low-frequency words. While Ha and Hyland (2017) and Chung and Nation (2003) do not include high-frequency technical terms without specialized meanings in their conceptualization at all, researchers such as Yang (1986) suggest a differentiation of technical and semi-technical terms. However, Liu and Lei (2020) dissuade from such a categorization, since some words are not easily placeable into one of these categories. Therefore, it is proposed that technical terms are regarded as unified concept including highly salient high-frequency, academic and low-frequency words with and without specialized meanings.

Regarding the identification of technical vocabulary two major methodological strands can be observed. Firstly, judgement based studies (e.g. Brieger & Phol, 2002; Chung & Nation, 2003) utilizing technical dictionaries (e.g. Chung & Nation, 2004; Ha & Hyland, 2017), rating scales (e.g. Chung & Nation, 2003) and/or contextual clues (Ha & Hyland, 2017) can be identified. Secondly, there are corpus-based studies (e.g. Bernier-Colborne & Drouin, 2014; Chung & Nation, 2003) that are based on computerized decisions using frequency and dispersion measures instead of more subjective human decisions (Liu & Lei, 2020). Additionally, many recent studies attempt to use a balanced approach incorporating features of both approaches (e.g. Ha & Hyland, 2017; Kwary, 2011; Watson Todd, 2017).

2.3.5 Academic vocabulary

While academic vocabulary is also associated with the terms “subtechnical vocabulary” (Anderson, 1980; Yang, 1986), “semitechnical vocabulary” (Farrell, 1990) and “specialized nontechnical lexis” (Cohen et al., 1988), Roessingh (2016) reports growing consensus regarding the importance of academic vocabulary, being a crucial aspect of EFL learners’ effort to assimilate to the academic discourse community (Coxhead, 2020). Nevertheless,

two main conceptualizations of English for academic purposes (EAP) vocabulary can be identified from the literature, namely English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP) (e.g. Coxhead, 2020; Hyland, 2016; Hyland & Tse, 2007). Advocates of EGAP (e.g. Coxhead, 2000; Gardner & Davies, 2014; Xue & Nation, 1984) share the perception that a shared core academic vocabulary, being “items which are reasonably frequent in a wide range of academic genres” (Hyland & Tse, 2007, p. 235) exists amongst all academic disciplines. In other words, this position assumes that EFL learners should acquire cross-disciplinary academic words and multiword units, suitable for all types of academic contexts (Coxhead, 2020; Dang, Coxhead & Webb, 2017). In contrast, the ESAP approach questions the existence and usefulness of such an interdisciplinary academic vocabulary and promotes researching and acquiring discipline-specific academic vocabulary (Hyland, 2016; Hyland & Tse, 2007).

While acknowledging the fact that terminology might be very specific for each academic discipline, this research study takes an EGAP approach, assuming that the discipline specific nature of vocabulary is already sufficiently covered by technical vocabulary lists, which incorporate not only technical low-frequency words exclusive to a discipline, but also include highly salient high-frequency, academic and low-frequency words. This is also supported by the fact that ESAP vocabulary lists partly overlap with vocabulary lists of technical vocabulary, being labelled ESAP vocabulary list or technical vocabulary list dependent on the position taken. For example, Chung and Nation’s (2003) vocabulary list of words specific to the medical field of anatomy is sometimes referred to as technical word list (Liu & Lei, 2020) but is also categorized as ESAP vocabulary list by others (Coxhead, 2018). Hence, this study adopts a similar theoretical position as found in Beck, McKeown and Kucan’s (2002) model of vocabulary knowledge, visualized in Figure 2.3 below.

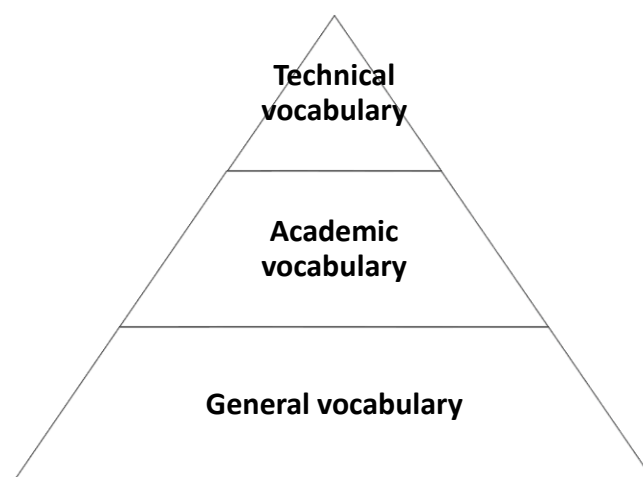


Figure 2.3: Beck, McKeown & Kucan’s model (2002) taken from Roessingh (2016)

As Beck, McKeown and Kucan (2002) exemplify in their model EFL learners first acquire general vocabulary knowledge, which comprises knowledge of high-frequency and mid-frequency words, followed by academic vocabulary (EGAP) and technical vocabulary specific

to their individual discipline and fields of interest. While agreeing to the differentiation of vocabulary knowledge in general vocabulary knowledge, consisting of high-frequency, mid-frequency and low-frequency word knowledge, academic vocabulary and technical vocabulary, the sequential nature of acquisition is questioned. This sequential nature is often attributed to the so-called *common core approach* in corpus studies, which assumes that knowledge of academic words builds on the knowledge of high-frequency words (Coxhead, 2018). Examples of *common core approaches* to academic vocabulary are the Academic Word List (AWL) of Coxhead (2000) or Browne et al.'s (2013) New Academic Word List (NAWL), which are built on top of existing high-frequency word lists, excluding members of these word lists. While this is often mentioned as a critical point of the AWL (Coxhead, 2000), which was developed on the basis of the GSL (West, 1953), including 570 highly frequent academic word families, it is proposed that the sequential nature might be laid out by the format but is not specifically suggested. Therefore, EFL learners still might acquire items of high-frequency lists simultaneously to those found in academic word lists. Hence, it is proposed that the existence of attuned high-frequency and academic vocabulary lists does not require successive acquisition of the words found on them but that both types of word lists might be used simultaneously and supportively.

Taking a closer look at EGAP vocabulary lists regardless of their method of compilation shows that several frequently used general academic word lists exist, such as the aforementioned *Academic Word List* (AWL) by Coxhead (2000) and Gardner and Davies (2014) *Academic Vocabulary List* (AVL). Moreover, Browne et al.'s (2013) *New Academic Word List* (NAWL) and Dang, Coxhead and Webbs' (2017) *Spoken Academic Word List* fall under the same category of academic word lists. The AWL (Coxhead, 2000) being the oldest of these lists encompasses the 570 most frequent academic word families across four disciplines, namely art, commerce, law and science (Coxhead & Bryd, 2012). Coxhead (2000) built her investigation on a corpus containing a total of 3 500 000 running words, nearly evenly distributed between the four disciplines. The 414 academic texts included academic textbooks, articles, book chapters and laboratory manuals adjusted for comparable length (Coxhead, 2000; Coxhead & Byrd, 2012). Over the years the AWL has shown a coverage between 11.60% (Cobb & Horst, 2004) to 8.96% (Coxhead & Hirsh, 2007) of the running words in university level texts, hence covering a substantial part of the running words of tertiary level academic texts. However, despite being the first and most successful EGAP word list, AWL (Coxhead, 2000) has been repeatedly criticized for various reasons, such as being based on the GSL and therefore potentially containing high-frequency words which are simply not in the GSL due to its outdatedness (Coxhead & Byrd, 2012; Gardner & Davies, 2007). Moreover, the counting unit of word families used has been questioned, since knowledge of word families requires high morphological ability in comparison to the

knowledge of lemmas¹ (Gardner & Davies, 2014). Regardless of these objections this research study is based on academic words extracted from the AWL, simply because the AWL is widely used and introduced in class to the participants of this study, which makes it more likely for students to show a development in their knowledge of academic words taken from the AWL than from any other more recent academic word list.

Similar to academic word lists of single words attempts have been made more recently to identify academic multiword units (Ackermann & Chen, 2013; Coxhead & Byrd, 2012; Durrant, 2009; Liu, 2012; Simpson-Vlach & Ellis, 2010). While most studies have compiled autonomous lists of academic multiword units, such as the Academic Collocation List (Ackermann & Chen, 2013) or Liu's (2012) list comprising the most frequent 228 academic multiword units, Coxhead and Byrd (2012) have discovered that collocations of items found in the AWL tend to occur in complex noun phrases or form collocations amongst each other, which has been affirmed by Ackermann and Chen (2013). On the basis of these findings it has been decided that this study will focus on verb and noun combinations of the target words taken from the AWL. This ensures that the knowledge of single word units and multiword units of the same target words are investigated. For more detailed information on the selection process of target words from the AWL see chapter 7.3.

2.4 Word knowledge

"If two people 'know' the meaning of a particular lexical item, their knowledge may not be identical" (Laufer et al., 2004, p. 205). This sentence perfectly introduces the concept of word knowledge, which might appear to be straight forward at first sight but is considerably more complex at second thought. Moreover, the wealth of research existent in the field of vocabulary learning has consequently led to an increase of terminology (Gyllstad, 2013). This chapter attempts to outline central concepts, dimensions and models proposed to describe word knowledge. Firstly, the difference between receptive and productive word knowledge will be addressed, followed by an introduction to the concepts of vocabulary size and vocabulary depth.

2.4.1 Productive and receptive knowledge

While several researchers (e.g. Cheng & Matthews, 2018; Meara, 1999; Melka, 2011) state that the process of finding a unified and consistent definition for receptive and productive word knowledge, also known as passive and active knowledge, has been difficult in the past, present researchers report a more transparent approach towards conceptualization (e.g. Cheng & Matthews, 2018; Henriksen, 1999; Miralpeix, 2020).

Daller, Milton and Treffers-Daller (2007, p. 6) amongst others (Fan, 2000; Miralpeix, 2020; Nation, 2013; Nation, 2019; Webb, 2008) state that receptively known lexical items are

¹ See chapter 2.1

“words you can handle in the context of reading or listening to speech”, while productive knowledge of a word means that “you can call [it] readily to mind when you need to speak or write [...]”. Hence, receptive word knowledge is needed for receptive skills and productive word knowledge likewise for productive language skills. In the case of receptive word knowledge this entails basic word knowledge such as the association of word meaning when encountering the word in a particular context, additional to more complex receptive abilities, such as recognising potential collocations of a word (Miralpeix, 2020). Similarly, productive word knowledge can range from being able to pronounce a word correctly in isolation to being able to use it correctly in a variety of contexts (Miralpeix, 2020). This range of receptive and productive word knowledge shades might be one of the potential reasons why the distinction of receptive and productive word knowledge has been criticized as being an “oversimplification” (Meara, 1999). At the same time, a study by González-Fernández and Schmitt (2019) shows that receptive and productive mastery seem to be statistically separate constructs. By using structural equation modelling (SEM) they conclude “that the distinction between [receptive] and [productive] knowledge is fundamental (González-Fernández & Schmitt, 2019, p. 481). However, several researchers reject the claim of receptive and productive knowledge being a strict dichotomy. Instead they propose that receptive and productive knowledge should be seen as a continuum (e.g. Cheng & Matthews, 2018; Henriksen, 1999; Melka, 2011), ranging from no knowledge over various degrees of receptive knowledge to complete productive mastery. This idea is partly based on the assumption that L2 acquisition partly aligns to L1 acquisition in so far that “with increased frequency of [receptive] exposure to a word comes increased richness of representation in the learner’s mental lexicon”, which leads to more productive control (e.g. Cheng & Matthews, 2018, p. 6; Henriksen, 1999). Hence, receptive knowledge is assumed to be acquired before productive knowledge (e.g. Cheng & Matthews, 2018; Daller, Milton & Treffers-Daller, 2007; Miralpeix, 2020). Indeed this suggestion is supported by the findings of several studies (e.g. Cheng & Matthews, 2018; Daller, Milton & Treffers-Daller, 2007; Miralpeix, 2020) showing that receptive vocabulary knowledge tends to be larger than productive vocabulary knowledge. However, this rule should not be overgeneralized, since individual characteristics, such as proficiency level, might have a considerable influence on the relationship of receptive and productive word knowledge (e.g. Daller, Milton & Treffers-Daller, 2007; Miralpeix, 2020).

To measure receptive and productive word knowledge with the help of vocabulary tests, Read (2000) suggests implementing two related terms, namely word recognition and word recall. To measure word recognition participants are confronted with test items, which present the words receptively to the learner and require the learner to demonstrate recognition or understanding of the words (Read, 2000). For instance, multiple choice formats as shown in example (1), hiding the target word amongst several distractors, can be considered to measure recognition knowledge.

(1) *leaf*

- a) an animal with feathers
- b) is growing on a tree
- c) a way of clothing in the middle ages

(2) The tree has got new green l_____.

In contrast, word recall requires the learners to actively produce the target word (Read, 2000). An example, would be a fill-in-the gap format, presenting a missing word in the context of a sentence as in example (2).

2.4.2 Vocabulary breadth and depth

While the study of word knowledge already dates back to the ancient Greeks and Romans (Milton & Fitzpatrick, 2014), Henriksen (p. 303) still emphasized the “need for more clarity and standardization” in 1999. A first step towards this goal was made by Anderson and Freebody in 1981 with their differentiation between vocabulary breadth and depth. In this widely accepted distinction (Staehr, 2009) vocabulary breadth is defined as “the number of words for which [a] person knows at least some of the significant aspects of meaning”, while vocabulary depth is known as “the quality or depth of understanding” (Anderson & Freebody, 1981, p. 93). Similar definitions were utilized more recently with Staehr (2009, p. 578) referring to vocabulary breadth as “the size of a learner’s vocabulary – that is, the number of words for which the learner has at least some knowledge of meaning” or Qian (2002, p. 515), who relates depth of word knowledge to “how well one knows a word”. Nevertheless, it remains unclear which type of word knowledge is exactly considered to fall under the categories of vocabulary breadth or vocabulary depth (Gyllstad, 2013). Researchers have attempted to answer this question from three different angles (Milton & Fitzpatrick, 2014; Laufer et al., 2004), namely from the so called component approach, the developmental approach and/or the metaphorical approach, which will be presented in the following. After introducing these main approaches their limitations and a new model of lexical knowledge, namely the ontogenesis model, will be discussed.

2.4.2.1 Component approach

The component approach is characterized by the fact that researchers following this tradition have attempted to provide lists of different word knowledge aspects. The earliest modern list of word knowledge components was proposed by Cronbach (1942), who distinguishes between knowledge of word definitions (generalization) and using them appropriately (application), multiple meaning knowledge (breadth of meaning), correct use of a word in different contexts (precision of meaning) and receptive or productive use (availability).

Richards (1976) expands on Cronbach’s (1942) early taxonomy by stating that word knowledge involves consideration of word frequency, register, syntax, derivation,

association, semantic features and polysemy, as nicely summarized by Qian (2002). However, it could be argued that even more word knowledge aspects are hidden in this early list, since knowing “the sort of words most likely to be found associated with the word” could also refer to collocational knowledge or “knowledge of the underlying form of a word” could be interpreted as knowledge of orthography. Moreover, different aspects of register related knowledge are named, with references to the mode in which a word is likely to be encountered and situational suitability of words, covering formality, currency or geographical restrictions. Hence, Richards’s (1976) list of word knowledge components can already be considered an extensive collection of word knowledge aspects, although not being very comprehensive or ordered. Laufer (1990) proposed another list of word knowledge components differentiating knowledge of form (phonological, graphic, morphological), syntactic behaviour, meaning (referential, associative, pragmatic) and relations to other words (paradigmatic, syntagmatic).

By incorporating aspects of word knowledge mentioned in earlier lists (Cronbach, 1942; Richards, 1976; Laufer, 1990), Nation (2013) produced “the latest and, to date, most comprehensive version of this type of analysis” (Milton & Fitzpatrick, 2014, p. 4), which is presented in his taxonomy of word knowledge depicted in Table 2.3 below.

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express 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?
	Concepts and Referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this word 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?

Table 2.3: Nation’s (2013) taxonomy of word knowledge

As visible, Nation (2013) differentiates between knowledge of word form, word meaning and word use, each encompassing three individual aspects of word knowledge, which in turn can be known receptively (R) and productively (P).

When taking a closer look at the category of word form, a learner can receptively recognize a lexical item when hearing or reading it, as indicated by receptive knowledge of the spoken or written form of a word. Productively, knowing the spoken form of a word would mean pronouncing a lexical item correctly, while productive written form means to adhere to spelling conventions (Nation, 2013). Moreover, a learner might recognize that a word is constructed out of several morphological units, such as suffixes or prefixes, which have particular functions and meanings (Nation, 2013; Nation, 2020). Hence, knowledge of word parts is mainly concerned with recognising and using the inflectional system of a given language correctly.

Similar to the knowledge of word form, the knowledge of word meaning is divided into three different aspects of word knowledge. Firstly, knowledge about word meaning includes the knowledge of the form-meaning link, which can be defined as the most basic connection between a word form and word meaning (Nation, 2013). The form-meaning link can be known both receptively and productively. Hence, being able to recognize a given word and coming up with its meaning or producing a particular word to express a particular meaning are considered to be the two dimensions of form-meaning knowledge (Nation, 2013; Nation, 2020). The form-meaning link is often considered to be one of “the most important aspects of vocabulary knowledge” (Nation, 2020, p. 18) alongside the knowledge of written and spoken word form (Gyllstad, 2013, p. 18). All three aspects of word knowledge have been used as basis to measure learners’ vocabulary size, since recognizing that a succession of letters or sounds make up a given written or spoken word form in a language or connecting a written word form to a meaning are considered to be the most central abilities of vocabulary knowledge (Laufer et al., 2004). In contrast, all other aspects of word knowledge presented in Nation’s (2013) taxonomy have been assigned to the concept of vocabulary depth (Gyllstad, 2013). While this study is not attempting to measure participants’ total vocabulary breadth, investigation of receptive and productive knowledge of the form-meaning link has been included due to its central position. Additionally, two more aspects are attempted to be researched, namely receptive collocational knowledge and receptive knowledge of multiple meanings, which are associated with the word knowledge aspect called concepts and referents.

The word knowledge component of concepts and referents is building on the knowledge of the form-meaning link and is concerned with multiple word meanings and the core meaning connecting all uses of a word (Nation, 2020). Exceptional cases are homophones, homographs and homonyms, sharing a particular word form but having two vastly different meanings (Nation, 2020). Again contextual word meaning can be known receptively by

recognizing multiple meanings of a word in different contexts or producing variable meaning senses related to context. Additionally, learners can recognize and produce associations to the target word, such as synonyms or antonyms, which is the last component of word meaning knowledge.

Regarding the use of words it is essential for learners to recognize, whether a given word is used grammatically correctly and to know how to use a word according to its grammatical functions (Nation, 2013). Furthermore, learners should be aware of potential collocational patterns in which a target word might occur and should be able to both recognize and produce them (Nation, 2013). Finally, learners need to be able to decide whether a word is suitable for a particular context or not. Nation (2020, p. 22) states that learners should have awareness of the level of formality, “politeness restrictions ([e.g.] fat, old, studied need to be avoided when talking to the person they apply to) [...]”, “geographical restrictions (dialects such as US English vs. UK English), age restrictions (language used to talk to children), datedness restrictions in that some words are old-fashioned and frequency restrictions”. Additionally, it is crucial to know whether a word is rather used in spoken or written mode of the language (Nation, 2020).

While this taxonomy is widely used amongst teachers and researchers alike, only few researchers (e.g. Chen & Truscott, 2010; Chui, 2006; González-Fernández & Schmitt, 2019; Schmitt, 1998; Webb, 2005; Webb, 2007;) have attempted to measure several word knowledge components at the same time. Hence, the multidimensional nature of vocabulary knowledge is often neglected, similar to the interrelationships between individual components (González-Fernández & Schmitt, 2019). The first researcher to investigate multiple word knowledge components was Schmitt (1998) when conducting his longitudinal study of the productive knowledge of written form, derivative knowledge, associative knowledge and multiple meaning knowledge of four academic EFL learners. He revealed that all four word knowledge aspects were related to each other and appeared to develop simultaneously (Schmitt, 1998). Similar to Schmitt (1998), Chui (2006) investigated meaning recall and productive derivational knowledge additionally to word class recognition and receptive collocational knowledge of 20 words from the AWL amongst 186 EFL learners. Moreover, Webb (2005; 2007) was also interested in the productive knowledge of the form-meaning link along with productive and receptive knowledge of written form, syntax, grammatical functions and associations. Finally, Chen and Truscott (2010) investigated the acquisition of orthography, grammatical function and associations, both receptively and productively, additionally to form-meaning recognition. Overall, the findings of these studies suggest that the individual word knowledge aspects are interrelated. However, it has to be considered that these studies do not cover all aspects of word knowledge, use different measurement methods, as well as different target words and target populations of participants. Therefore, the multicomponent nature of vocabulary knowledge can be considered a research gap waiting to be filled by future studies.

Through the investigation of three different word knowledge aspects, namely form-meaning link knowledge (receptive and productive), multiple meaning knowledge (receptive) and collocational knowledge (receptive) of academic verbs this research adds to this growing field of vocabulary knowledge. By performing statistical measures, such as correlations and repeated measures ANOVAs, an insight into the relationship between the three components can be given.

2.4.2.2 Developmental approach

The developmental approach to word knowledge is building on the component approach and investigates whether an order of acquisition of word knowledge aspects can be found (Milton & Fitzpatrick, 2014). Hence, the developmental approach is interested in the fact whether some aspects of word knowledge are preceded by other components when learning a language. Several researchers (e.g. Dale, 1965; Paribakht & Wesche, 1993) have tried to place individual steps in the acquisition of word knowledge on a continuum ranging from no word knowledge to complete mastery of a word.

An early attempt was made by Dale (1965), who suggested that spoken and written form recognition proceeds knowledge of the form-meaning link by proposing the following model:

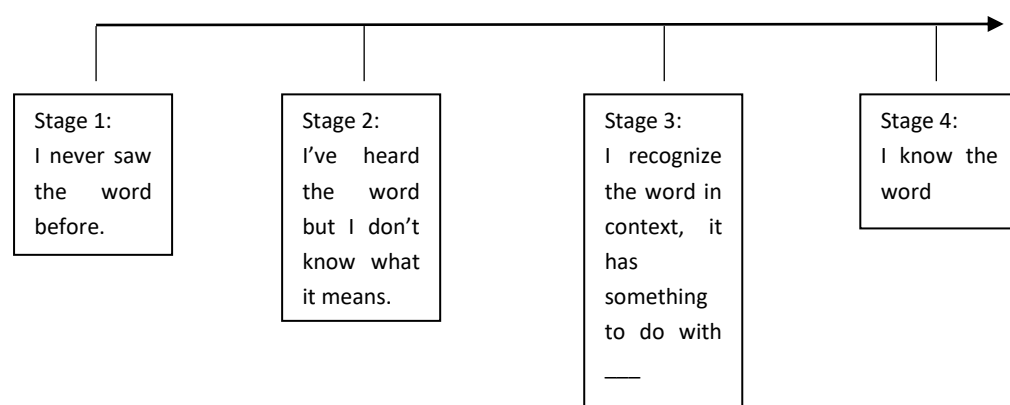


Figure 2.4: Developmental model by Dale (1965)

The most important assumption made by Dale (1965) is that form recognition proceeds knowledge of word meaning, which was adopted by Paribakht and Wesche (1993) in their widely used Vocabulary Knowledge Scale (VKS). This hierarchical model comprises five stages instead of four as proposed by Dale (1965). As visible below Stage 1 corresponds to no word knowledge, while stages two and three are comparable to Dale's (1965) interpretation, considering receptive knowledge of written and spoken form and the form-meaning link.

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)

However, Paribakht and Wesche (1993) additionally include aspects of word use, such as knowledge of constraints on use and grammatical function. Therefore, they propose that the order of acquisition of different word knowledge components resembles Nation's (2013) taxonomy of word knowledge in so far that knowledge of word form proceeds knowledge of word meaning, which in turn is necessary for knowledge of word use.

While developmental scales, such as the VKS, are in use, most of these models have not been empirically tested. Instead these assumptions are purely theoretically driven in most cases. One of the rare exceptions to be named is Chui (2006), who investigated the order of acquisition of four word knowledge components, namely word class recognition, meaning recall, collocation recognition and derivative form recall of 186 EFL learners. She reports that students performed better on word class recognition and meaning recall, which might indicate that these aspects of word knowledge are acquired earlier than collocation recognition and derivative form production (Chui, 2006). However, this conclusion is only based on one data collection, without any observation of development within a period of investigation. In contrast, Webb (2005; 2007) has examined which of five components of word knowledge has the highest gains when asking learners to learn non-words over a period of time. He found that orthography showed the steepest development while the other components, consisting of receptive and productive knowledge of the form-meaning link, syntagmatic and paradigmatic associations and grammatical functions, could not be consistently placed over the period of investigation (Webb, 2005; Webb, 2007). Hence, Webb's (2005; 2007) studies suggest that the progress in individual areas of word knowledge might increase and/or fluctuate over a learning period, creating different developmental orderings. However, evidence for and information about a potential acquisition order is rather thin, which is why González-Fernández and Schmitt (2019) undertook the first empirical study based on complex statistics to investigate the order of acquisition of receptive and productive form-meaning link knowledge, collocation knowledge, multiple meaning knowledge and derivative knowledge. By using implicational scaling (Guttman, 1944) and a Mokken analysis (Molenaar & Sijtsma, 2000) they arrived at the following order of acquisition for their data of 144 Spanish speaking EFL learners of various proficiency levels:

Form-meaning link recognition > Collocate form recognition > Multiple-meaning recognition > Derivative form recognition > Collocate form recall > Form-meaning link form recall > Derivative form recall > Multiple-meaning recall

González-Fernández & Schmitt (2019, p. 493)

This scale implies that several word knowledge components need to be known receptively before being produced. This study is going to attempt a similar analysis of three word knowledge aspects investigated, which will show whether the acquisition order found in González-Fernández and Schmitt (2019) is also stable for another data sample.

2.4.2.3 Metaphorical approach

Simultaneously to the component approach and the developmental approach several researchers (e.g. Aitchison, 1987; Daller, Milton & Treffers-Daller, 2007; Schur, 2007) have worked on a metaphorical approach to word knowledge. One well known metaphor is the idea of a web of words, which depicts potential connections between individual words as visible in Figure 2.5.

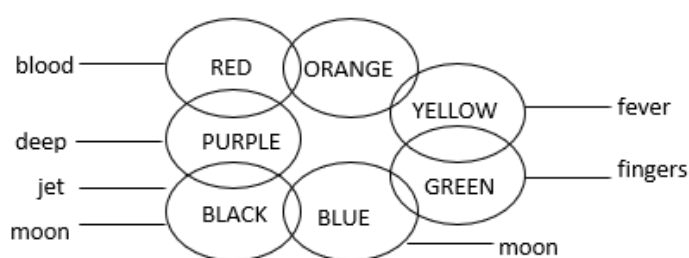


Figure 2.5: Word web (Aitchison, 2004, p. 94)

This graphical representation, which was originally suggested by Aitchison (1987, p. 84), is frequently used to picture the dimension of vocabulary use and has also been of empirical use (e.g. Meara & Wolter, 2004; Schur, 2007).

A similarly fruitful metaphor is that of lexical space (Daller, Milton & Treffers-Daller, 2007), which includes the dimensions of vocabulary depth, breadth and fluency (see Figure 2.6).

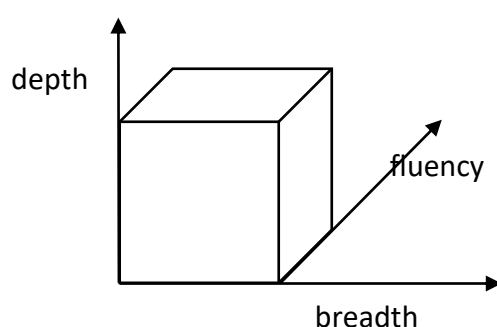


Figure 2.6: The lexical space (see Daller, Milton & Treffers-Daller, 2007, p. 8)

While the axis of vocabulary breadth and depth have already been introduced before, the third axis representing fluency needs further explanation. Daller, Milton and Treffers-Daller (2007, p. 8) describe fluency as “how readily and automatically a learner is able to use the words they know and the information they have on the use of these words”. Hence, fluency comprises aspects such as speed and accuracy of receptive and productive word use (Daller, Milton & Treffers-Daller, 2007). As Milton and Fitzpatrick (2014) highlight this metaphor

allows a placement of learners with different word knowledge profiles in the three-dimensional space. However, this study does not include the dimension of fluency but is restricted to measurements of knowledge of the form-meaning link and other aspects of vocabulary depth.

As visible in the prior chapters several attempts to describe and define word knowledge have been made and different approaches have led to different dimensions, constructs and models. However, it has to be noted that many of these models have overlaps and communalities. For instance, Gyllstad (2013) outlines that Meara's (2005) model of lexical competence/skill shares similarities with Daller, Milton and Treffers-Daller's (2007) model of lexical space and Henriksen's (1999) model of lexical development. Hence, the "conceptual overlap" between competing models/attempts should not be neglected similar to different operationalizations proposed (Gyllstad, 2013, p. 14).

2.4.2.4 Challenges of vocabulary breadth and depth

All approaches outlined above classically assume the differentiation of vocabulary breadth, consisting of knowledge of the written and spoken form of a word and the form-meaning link, and vocabulary depth, subsuming all other categories of Nation's (2013) taxonomy. However, neither the basic conceptualization of vocabulary breadth and depth as opposite poles, nor their exact definitions have been firmly established on empirical grounds. On the contrary, first empirical findings give reasons for a potential rethinking of these dimensions.

By using structural equation modelling (SEM) González-Fernández and Schmitt (2019) tested whether Nation's (2013) conceptualization of word knowledge as depicted in Figure 2.7 below, was suitable for their data of 144 Spanish speaking EFL learners of various ages and proficiency levels.

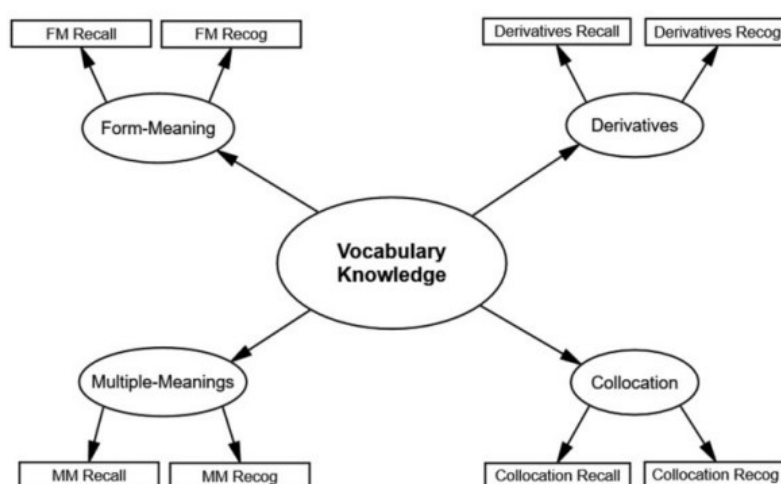


Figure 2.7: Hypothesized model by González-Fernández & Schmitt (2019)

They tested the receptive and productive knowledge of four different components, namely knowledge of the form-meaning link, derivational knowledge, knowledge of multiple

meanings and collocational knowledge. As visible in Figure 2.7 they hypothesised that the receptive and productive side of each component measured would add to an overarching word knowledge component, which in turn adds to the main concept of vocabulary knowledge, as implied by Nation's (2013) taxonomy. However, SEM showed that, although a good model fit was given, the four components were too highly intercorrelated, indicating that they should be "understood as single construct" (González-Fernández & Schmitt, 2019, p. 494). Therefore, the model in Figure 2.7 based on Nation's taxonomy does not appear to be suitable for the collected data (González-Fernández & Schmitt, 2019). Instead González-Fernández and Schmitt (2019, p. 495) conceived another model, which considers "recognition and recall aspects as individual direct indicators of the general vocabulary knowledge construct".

While few findings have been reported yet, these first results also seem to be opposed to the sharp division into vocabulary size and vocabulary depth, since concepts underlying both dimensions have been included in the model but seem to be closely related. Other studies (e.g. Nurweni & Read, 1999; Meara & Wolter, 2004; Gyllstad, 2007) using correlations further support these findings. This suggests that vocabulary breadth and vocabulary depth should not be conceptualized as opposites.

However, additional to the unclear relation of vocabulary breadth and vocabulary depth, both concepts exhibit vague or challengeable definitions. Regarding the definition of vocabulary breadth, Gyllstad (2013) calls the attention to the role of multiword units for the dimension of vocabulary breadth. As Shin and Nation (2008) have shown multiword units, such as collocations are used with such high-frequency that many of them would qualify to be included in high-frequency vocabulary lists, challenging the classical definition of vocabulary size and their original placement in the depth dimension (Gyllstad, 2013, p. 22).

Moreover, taking a closer look at Nation's (2013) taxonomy and the dimensions traditionally assigned to vocabulary depth shows that the nature of vocabulary depth appears to be highly multi-faceted, making operationalization challenging (Gyllstad, 2013, p. 22).

2.4.2.5 The ontogenesis model

A very recent model proposed by Bordag, Gor and Opitz (2022), called the ontogenesis model, offers an innovative approach to the conceptualization of lexical knowledge and its development, bringing together several research strands on lexical processing and vocabulary learning (Gyllstad, 2022). Consequently, it will be seen in how far the results of this study can be linked to the ontogenesis model.

A basic feature of the model is that it tries to map lexical development from the emergence of a lexical representation, its way to the optimum of lexical representation, attrition and forgetting. To fulfil this aim Bordag, Gor and Opitz (2022) make use of two basic concepts, optimum and fuzziness. They state that at its optima a lexical representation is "properly

encoded and fully specified” (Bordag, Gor & Opitz, 2022, p. 185). Hence, the peak of lexical knowledge for this lexical representation has been reached. Corresponding to this the way from the emergence of a lexical representation to it reaching its optimum is characterised by fuzziness. According to Bordag, Gor and Opitz (2022, p. 186) “[f]uzziness refers to inexact or ambiguous encoding of different components or dimensions of the lexical representation [...]”. Hence, fuzziness can be described as partial knowledge of the lexical representation, which is an intrinsic characteristic of the mental lexicon. In Figure 2.8 below Bordag, Gor & Opitz (2022, p. 187) visualize the concepts of fuzziness and optimum.

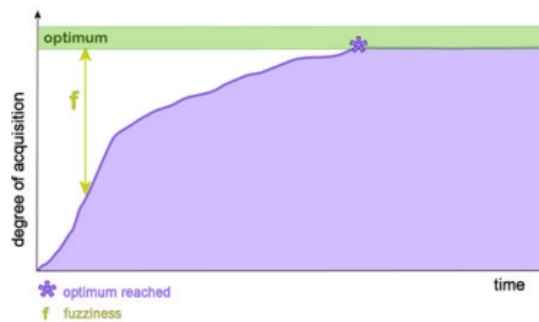


Figure 2.8: Illustration of the concepts of fuzziness and optimum taken from Bordag, Gor & Opitz (2022, p. 187)

In addition to these fundamental assumptions the ontogenesis model is multidimensional, comprising three main dimensions (Bordag, Gor & Opitz, 2022). The first dimension is the dimension of linguistic domains, which builds on research on vocabulary depth, meaning that linguistic domains “correspond to the various types of word knowledge that need to be represented” (Bordag, Gor & Opitz, 2022, p. 186). The recent version of the ontogenesis model focuses on semantic, phonological and orthographic components, which are assumed to develop gradually for each component. Figure 2.9 below shows the development of all three domains based on the concepts of fuzziness and optima.

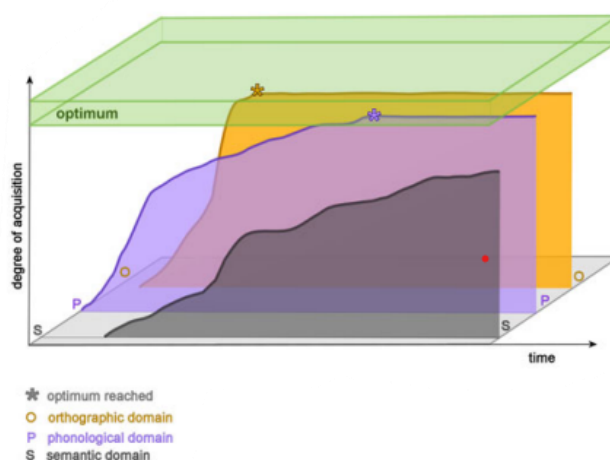


Figure 2.9: Illustration of the dimension of linguistic domains taken from Bordag, Gor & Opitz (2022, p. 187)

As visible not all three domains emerge at the same time or proceed towards the optima at the same pace. For instance, in Figure 2.9 the phonological domain of the lexical representation emerges first, followed by the orthographic domain and lastly the semantic domain. However, the orthographic domain displays the steepest development and reaches the optimum first, followed by the phonological domain. In contrast, the semantic domain does not reach its optimum at all. Therefore, the ontogenesis model can be understood as an extension of existent models, such as Wolter (2001), which have determined that words develop individually, since it proposes that not only words but also their components are developing individually (Bordag, Gor & Opitz, 2022). However, this should not be misunderstood as a claim that the individual domains are unrelated. On the contrary, Bordag, Gor and Opitz (2022, p. 187) specifically state that the developmental curves of the individual domains are related and that components are developed “gradually and cumulatively”. Hence, the development can be considered to be an ontogenetic process.

While the dimension of linguistic domains has already built on what has been outlined in the component approach before, the second dimension of the model, namely the dimension of mappings, points at a research gap in the field of vocabulary research. While the first dimension is concerned with the parallel development of the individual components, the second dimension considers how well the knowledge of these domains is linked on a continuum ranging from weak to robust connections (Bordag, Gor & Opitz, 2022). These connections are referred to as mappings, which are closer defined as “connections that are activated to bond the identities of words in memory” (Ehri, 2004). Figure 2.10 below visualizes this dimension exemplarily with the phonological and the semantic domain.

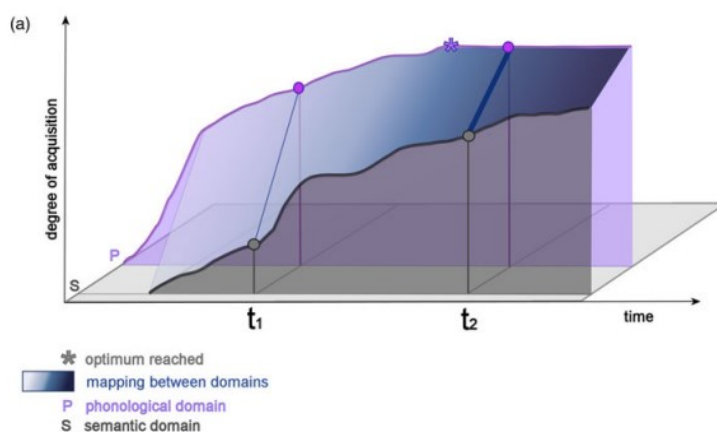


Figure 2.10: Illustration of the dimension of mappings taken from Bordag, Gor & Opitz (2022, p. 188)

Taking a closer look at the illustration shows that the strength of the connections between the two domains is depicted at two points in time, t_1 and t_2 . While the line between the two components is still very thin at t_1 the connection has grown stronger at t_2 . The development of the dimension of mappings is also visualized with the help of the colour continuum ranging from nearly transparent over light blue to dark blue. The suggestion underlying this

dimension is that the stronger the mappings are the easier the retrieval of the lexical representation should be. Hence, this dimension can be partly connected to the concept of fluency in the metaphor of lexical space (Daller, Milton & Treffers-Daller, 2007), since one could assume that with the ease of retrieval the fluency of a person should increase.

Lastly, Bordag, Gor and Opitz (2022) integrate the dimension of networks into the ontogenesis model. This dimension is concerned with “how a given lexical representation is interconnected with other units in the given language” (Bordag, Gor & Opitz, 2022, p. 187). A lexical representation can only be connected scarcely with other units and these connections can also vary in strength. Therefore, the dimension of networks comprises many of the aspects associated with word use (Nation, 2013), such as association or collocation knowledge. Figure 2.11 exemplarily shows the developmental curve of one domain (e.g. phonological domain) over time.

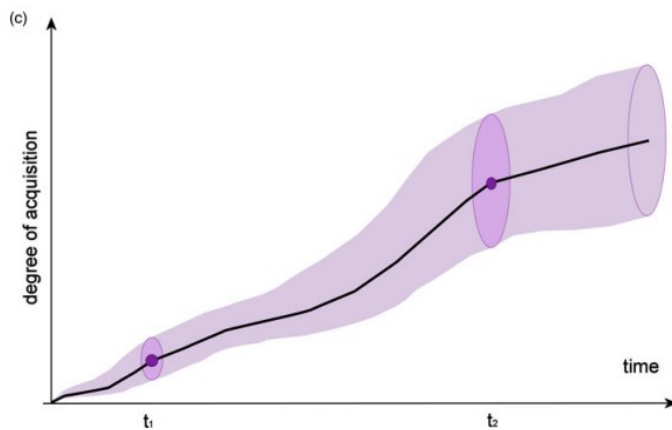


Figure 2.11: Illustration of the dimension of networks taken from Bordag, Gor & Opitz (2022, p. 189)

“Depicted over time, the circles representing network integration yield a cone-like structure around the curve in the three-dimensional space” (Bordag, Gor & Opitz, 2022, p. 189). Two points in time, t_1 and t_2 are highlighted, which show that the radius grows when the lexical representations become more integrated in the network, meaning that they have more and stronger connections with other units in the language.

Drawn together these three dimensions result in the following model of the non-native lexicon for the phonological and the semantic domain:

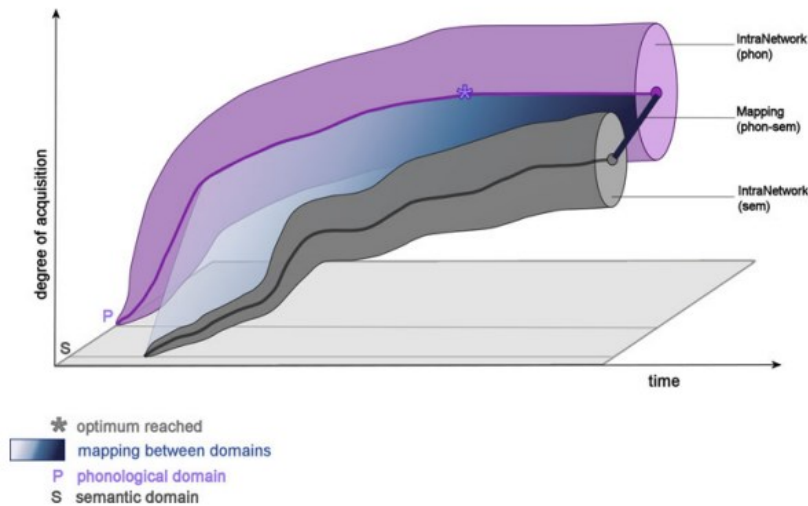


Figure 2.12: Illustration of the ontogenesis model for the phonological and semantic domain taken from Bordag, Gor & Opitz (2022, p. 191)

However, Bordag, Gor and Opitz (2022) additionally propose that the model can also be used for L1 speakers and that most lexical representations reach their optimum and stay there for an L1 speaker, which is not the case for L2 speakers.

While the ontogenesis model can be seen as a fruitful innovative approach uniting several parallel research fields and areas, it should not be considered to be concluded yet (Gyllstad, 2022). Instead Gyllstad (2022, p. 216) suggests to use it “as a point of departure for further refinement”. For example, Gyllstad (2022) outlines two points for future clarification. Firstly, he names the representation of multiword units in the mental lexicon either as individual lexical representation putting down roots or as holistic concept processed as a whole (Gyllstad, 2022). Secondly, he indicates some unclear points regarding the concepts of optimum and fuzziness in L1 and L2 lexicon (Gyllstad, 2022). Moreover, the model opens up several questions with regard to the integration of prior research conducted. Firstly, the question arises how the very prominent differentiation of receptive and productive or recognition or recall knowledge for different components are mirrored in the ontogenesis model, since the authors do not elaborate on these concept in much detail. When outlining the relation to the Complementary Learning Systems (CLS) account (Lindsay & Gaskell, 2010), the authors state:

According to the account, an episodic memory trace is formed first (supported by the medial temporal lobe), which is sufficient for explicit recognition and recall of novel lexical items (corresponds to the domain dimension in the OM) (Bordag, Gor & Opitz, 2022, p. 190).

Hence, it can be assumed that both dimensions, recognition and recall knowledge of each component are reflected in its ontogenetic curve starting from emergence to the optimum (and eventually loss through attrition). This would be supported by research suggesting that

recognition knowledge of a word usually is easier than recall knowledge (e.g. González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004). However, it still remains unclear which processes underlie this transition from recognition to recall knowledge and which factors interact with this development. At the same time, it is an intriguing thought that the readiness of a lexical representation to be used receptively to productively is influenced by the degree of fuzziness existent for the entry at the given time as well as by the strength of mappings between components and could be mapped in an ontogenetic curve.

In addition, a closer look at Nation's (2013) taxonomy and how it is represented in the ontogenesis model reveals that some aspects still might require further information. While the word knowledge components spoken and written form have been selected as two of the main three linguistic domains, namely phonology and orthography, based on prior research by Perfetti & Hart (2001, 2002) and Perfetti (2007), the third aspect of word form knowledge, namely knowledge of word parts is not explicitly placed in the model. The same accounts for grammatical knowledge found in the category of word use. However, Bordag, Gor & Opitz (2022, p. 186) reveal in a footnote that an extra paper will be devoted to this gap regarding the role of morpho-syntax in the ontogenesis model in the future.

Moving on the third linguistic domain considered in the model, namely semantics, it reveals that this domain partly subsumes two components in Nation's (2013) taxonomy, namely knowledge of the form-meaning link and concepts and referents, subsumed as "all core senses of the word" by Bordag, Gor & Opitz (2022, p. 188). Nonetheless, it still remains rather open where the word knowledge aspect on constraints on use in Nation's (2013) taxonomy is placed in the model. Taking a look at Bordag, Gor & Opitz's (2022, p. 188) statement about the fact that "the semantic optimum may differ to some degree for individual speakers (e.g., specific senses may be available only to experts in particular fields)", implies that this component might be integrated in the semantic domain, being part of the ontogenetic curve of the domain. However, since this component subsumes much more knowledge about the lexical representation, such as the level of formality, the currency or predominant occurrence in written or spoken mode, a closer definition would be necessary. Nevertheless, it has to be recorded that the lack of a complete and definite definition of the component constraints on use was already inherent in Nation's (2013) taxonomy.

Linked to the representation of the different word knowledge aspects in the ontogenesis model is the presence and definition of the dimensions of vocabulary breadth and depth. While prior research has widely followed the assumption that orthographic, phonological and basic semantic knowledge belong to the dimension of vocabulary breadth, the ontogenesis model has chosen exactly these dimensions as pillars for their dimension of linguistic domains, depicting their development from emergence to optimum, similar to the basic idea underlying the classical concept of vocabulary depth. Although the authors do not

explicitly mention the concept of vocabulary breadth in relation to their model, it might be assumed that they align with the most basic definition of vocabulary breadth, being the sheer number of words known. Hence, a representation similar to the one in Figure 2.12 should be existent for each lexical entry in the mental lexicon. However, a solution for the connection between these individual ontogenetic models per word needs to be developed.

Finally, a stronger emphasis on the fact that the word form can but does not HAVE to be the first aspect encountered is suggested. For example, a learner pursuing a receptive activity such as extensive reading, might encounter an unknown and new orthographic form and might then link this orthographic form to a meaning by use of different strategies, such as inferencing or use of a dictionary (see Figure 2.13 on the left).

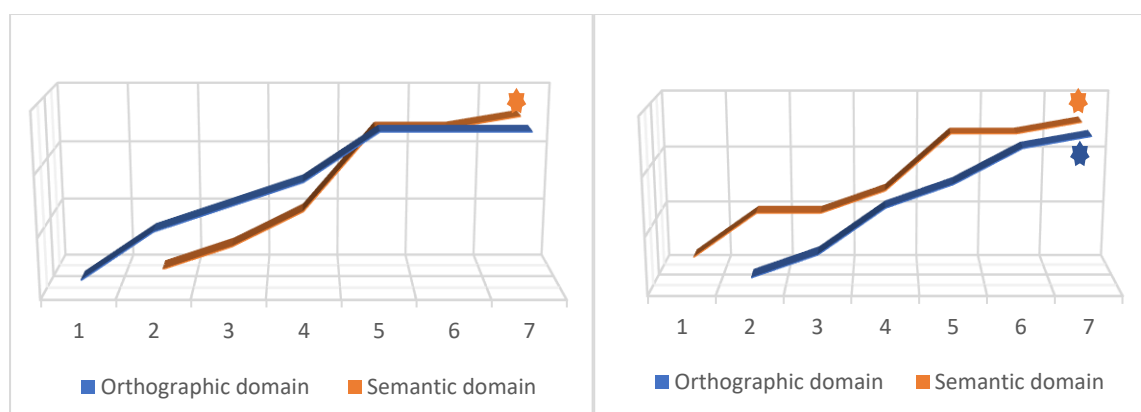


Figure 2.13: Illustration of primary domain for two activities

However, another situation, such as a productive activity as writing an L2 test, might leave the learner searching for an orthographic form in the L2 that corresponds to an already existent L1 meaning (see Figure 2.13 on the right). Hence, the lexical learning process would start with a semantic concept for which an orthographic form is learned. Of course there is no guarantee that the L1 semantic concept in the learners head and the L2 semantic concept related to the word form found are 100% congruent, as outlined by Bordag, Gor and Opitz (2022). However, this would then be reflected in the degree of fuzziness.

3 Vocabulary testing

Since the theoretical dimensions of vocabulary knowledge, outlined in chapter 2, are not directly observable, they need to be made visible and measurable, which is frequently attempted by the use of vocabulary tests (e.g. Qian, 2002; Schmitt, Schmitt & Clapham, 2001). Hence, this chapter will first provide an overview of the most important concepts in the area of vocabulary testing, namely the definition of a test construct, test reliability and test validity. This will be followed by a critical introduction to a selection of existent vocabulary tests, measuring comparable constructs as in this study.

3.1 Testing theory

As many researchers have noted critically in recent studies (e.g. Daniel & Kremmel, 2020; Kremmel, 2016; Kremmel & Harding, 2020; Kremmel & Schmitt, 2016; Schmitt, Nation & Kremmel, 2020) many existent vocabulary tests lack consistent development criteria and validation evidence. Hence, this chapter will present the three most crucial quality criteria for tests, namely construct definition, reliability and validity and will relate them to the tests used in this research study.

3.1.1 Construct

As Gyllstad (2007; 2013, p. 12) outlines, the term construct originates from the field of psychology but has come into use in language testing, being “a form of knowledge or ability that can be observed and/or measured”. Hence, a construct describes what a test measures and what test scores are able to tell us about the test candidates. Bachman (1990) differentiates between three stages of construct definition for test development:

- a) The construct needs to be defined theoretically.
- b) The construct needs to be defined operationally.
- c) Procedures must be established for the quantification of observations.

(Gyllstad, 2007, p. 33)

This conceptualization reflects the interactionist view of test construct definition (Chapelle, 1998), fusing the trait and the behaviourist approach. While trait theorists “would attribute test performance [solely] to the characteristics of the learner” (Chapelle, 1998, p. 35), behaviourists “would include the features of context relevant to vocabulary use” (Chapelle, 1998, p. 38). Hence, trait theorists interpret test scores as a direct indication of learners underlying abilities, without acknowledgment of contextual surroundings, while behaviourists attribute most of the test scores to different environmental and sociolinguistic contexts instead of viewing them as representations of learner traits. The interactionist perspective on the definition of a test construct can be seen as a fusion of both worlds, resulting in test performance being defined as “a sign of underlying traits [which is] [...] influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts” (Chapelle, 1998, p. 43). Endorsing Chapelle’s (1998) preference of the interactionist perspective on construct definition, this study adheres to these norms.

Whereas a theoretical definition of the constructs measured by the tests developed and used in this study has already been attempted in chapter 2, which has thematized the construct of vocabulary knowledge in more detail, the operationalization of the constructs to be measured is mainly addressed in chapter 6. Moreover, the following chapter 3.1.2. will provide more insight into common ways of operationalizing comparable constructs and the results yielded.

Regarding the stage of quantification listed in Bachman's (1990) compendium of construct definition each of the four word knowledge components measured is analyzed with the help of quantitative measures. Each of the tests constructed results in an interval scale (e.g. Field, 2018), which allows in depth analysis of the individual variables, as well as their relationships to each other.

3.1.2 Reliability

Apart from the necessity to clarify the construct definition of each test used, each test should be tested for its reliability before being applied. As the everyday usage of the word reliability suggests, test reliability can be defined as "the accuracy with which a test measures what it is supposed to measure" (Daller, Milton & Treffers-Daller, 2007, p. 15). Hence, if researchers want to investigate the reliability of a test, they want to examine how consistently this test measures the underlying construct. However, a test score can be caused by other factors apart from the underlying construct to be measured. Bachman (1990, p. 350) proposes that variance in the scores of a language test can be attributed to four categories:

- a) Language ability
- b) Personal characteristics
- c) Test method
- d) Random factors

While the language ability can be referred to as the "true score component", being the underlying ability tested, personal characteristics, test methods and random factors can be summarized as "error score component" (Gyllstad, 2007, p. 65), which involuntarily influence the test score. Personal characteristics, being biographical variables such as gender or age, and the test method used are further referred to as systematic factors, "since two individuals who differ in terms of these factors will perform differently on a test (Bachman, 2004, p. 156; Gyllstad, 2007, p. 64). Conversely, random factors are unsystematic, since they include differences in test administration or in the daily condition of test takers, which might or might not influence the test score. Hence, the observed score (x) of a test consists of the sum of the true score (x_t), caused by the true score component and the error score (x_e), caused by the three error score components (Bachmann, 2004, p. 158):

$$x = x_t + x_e$$

Hence, test reliability is the proportion of the observed score caused by the true score or as Gyllstad (2007, p. 63) formulates it: "Reliability is the absence of measurement error". This means that the less measurement error is caused by error score components the higher the reliability of a test is.

After clarifying what reliability is, the question remains how reliability might actually be measured. Lado's (1961, p. 330) definition of reliability allows more insight into how this might be attempted:

Reliability has to do with the stability of scores for the same individuals. If the scores of students are stable the test is reliable; if the scores tend to fluctuate for no apparent reason, the test is unreliable.

As inferred by this definition, reliability can be investigated by asking a test taker to take the same test twice without any changes in their knowledge (Daller, Milton & Treffers-Daller, 2007; Gyllstad, 2007, p. 63). If the scores on the first and the second sitting of the test remain stable for each test taker the test can be considered to be reliable. A more general example would be, if we measure the height of a group of children at 10 am and again at 10:30 am, we would expect the children to be of the same height at both points in time. If not, we would probably start to question the reliability of the measurement instrument, in this case a tape measure. The same principle applies to other measurement instruments, such as tests. This type of reliability is often referred to as test-retest reliability (Field, 2018, p. 822). However, the test-retest procedure has several flaws, such as the possibility of restricted availability of the participants for two sittings and potential practice effects, which might lead to an increase of scores due to the fact that the test has already been taken a short time before the second sitting. Therefore, the present thesis uses a reliability coefficient of internal consistency, which allows a calculation of reliability from one score only (Gyllstad, 2007, p. 65; Hughes, 2020). Instead of calculating correlation coefficients for the scores of a first and a second sitting, reliability coefficients of internal consistency are achieved by splitting the data of one data collection in every mathematically possible manner and calculating the correlation coefficients for each of these splits. For these correlation coefficients an average is computed, which is known as the Cronbach's alpha (Cronbach, 1951). The formula for Cronbach's coefficient α is given in Figure 3.1 below:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s^2 i}{s^2 x} \right)$$

Figure 3.1: Formula for Cronbach's coefficient α (Bachmann, 2004, p. 163)

Note: K=number of items in the test; $\sum s^2 i$ = sum of item variances; $s^2 x$ = variance of the test scores

In general, a Cronbach's α of 0.70-0.80 is considered to be acceptable and a value of 0.85-0.90 desirable (Brown, 1983; Field, 2018; Gyllstad, 2007). However, Brown (1983) lists four aspects, which might influence the Cronbach's α score of a test:

- a) Test length
- b) Range of scores
- c) Test difficulty
- d) Too low time limit

Regarding test length it has been observed that the more items a test contains the more reliable it gets, which can be explained by the fact that random measurement errors, as mentioned above, balance each other. For instance, errors caused by test fatigue will rule out points achieved by guessing (Gyllstad, 2007, p. 66). Similarly, the range of scores obtained for a test influences the reliability coefficient. While homogenous groups display a low variance, which decreases the reliability score, heterogeneous groups with a large range, will have larger reliability scores. This is directly related to issues of test difficulty, since if a test is too easy, a so-called ceiling effect (Davies et al., 1999) is caused and if a test is too difficult, a floor effect is caused, both leading to a small range of scores and a decrease of reliability. Finally, reliability scores for speeded tests need to be treated with caution, since test takers might have had too little time to complete the test, which will result in missing answers disturbing the reliability score.

Finally, Gyllstad (2007, p. 66-67) emphasizes that a reliability score is always related to the test scores and not to the test itself. This means that one and the same test might behave very differently, if used with different target groups (Gyllstad, 2019). Hence, it is of crucial importance to pilot a test with a group similar to the target group to test its reliability behavior.

3.1.3 Validity

In addition to the necessity to investigate a test's reliability, every test developer should reflect whether their interpretation of performance based on the results of the designed tests are well justified, which is referred to as the validity of a test (Chapelle, 1998, p. 48; Gyllstad, 2007; Henning, 1987). Messick (1989, p. 13) broadly defines validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores". However, the validity can also be seen as consisting of various subcomponents, instead of being a unitary concept (Gyllstad, 2007, p. 67). Many researchers, such as Henning (1987), Davies et al. (1999), Bachman (1990), Brown (1983), Messick (1989) or Hughes (2020) differentiate the following types of validity:

- a) Construct validity
- b) Concurrent validity
- c) Predictive validity
- d) Content validity
- e) Response validity
- f) Face validity

While construct validity, concurrent validity and predictive validity can be investigated both empirically and logically, content validity, response validity and face validity are primarily established on theoretical grounds. Chapelle (1998, p. 50) defines construct validity as "judgmental and empirical justifications supporting the inferences made from test scores".

In other words construct validity tries to manifest whether the test performance can be interpreted as indicator of the underlying construct that should be investigated. This can be achieved by measures of internal consistency, such as the Cronbach α coefficient, treated in the section on reliability, or by experimental research which investigates whether the performance on the test differs across groups as expected (Chapelle, 1998; Gyllstad, 2007, p. 68). For instance, it can be expected that participants with a lower overall proficiency will have a larger gap between their receptive and productive form-meaning link knowledge of words than more advanced language learners. If a newly developed test confirms this expectation from prior studies, its construct validity is strengthened. This research study has examined the internal consistency of the measured constructs based on the Cronbach α coefficient, as well as on explorative factor analyses, which will be outlined in more detail in the methodology section.

Regarding concurrent validity Chapelle (1998, p. 52) mentions that this type of validity of a test can be examined by comparing the results of a new test to an already established test of the field, which is supposed to measure the same construct. If a high correlation is found between the new and the established test, the concurrent or criterion-related validity of the test can be considered to be high (Henning, 1987; Hughes, 2020, p.31). However, it has to be mentioned that while established and widely used tests of receptive and productive form-meaning link knowledge as well as receptive collocation knowledge exist in the field of vocabulary research, most of these tests have been criticized due to their lack of validation. Moreover, none of these tests have been developed specifically for a comparable target group or cover similar target items as the tests developed in this study, which makes them only comparable to a limited extent. Hence, this study does not investigate concurrent validity in the traditional sense but has instead relied on empirical item investigation, such as the examination of item difficulty and item discrimination.

Predictive validity is highly related to concurrent validity in the sense that it is concerned with relationships between the abilities tested and other proficiencies. However, “the focus of ‘predictive validity’ is the examination of whether scores on a test may predict future language behavior” (Gyllstad, 2007, p. 69).

In contrast content validity is concerned with the question whether the test covers suitable content and whether it is relevant for the language ability under investigation (Gyllstad, 2007; Hughes, 2020, p. 29). This can either be examined theoretically through thorough literature research of the given field, or with the help of expert judgments (Hughes, 2020, p. 30). The latter scenario requires experts trained in the research field in question to judge what they believe a test is measuring. If the analysis of the experts is close to what the test is intended to measure, this can be counted as evidence for a test’s content validity. This study is based on a thorough literature review of all research fields involved and advice from

experts has also been gathered throughout the test development process, increasing the content validity of the respective tests.

Furthermore, response validity is concerned with the extent to which test takers respond in the manner expected by the test developers (Henning, 1987, p. 92). To increase response validity test instructions and test format can be examined for their clarity (Gyllstad, 2007, p. 69). This could be investigated by implementing think aloud protocols while taking the test, by asking participants to indicate problems or by asking them to unveil guessing (Chapelle, 1998). The latter option has been chosen in this research study.

Lastly, face validity being “the extent to which a test measures what it is supposed to measure in the eyes of [...] the test takers”, has been investigated by inviting participants’ feedback on the tests after taking them (Gyllstad, 2007, p. 69).

While consideration of multiple layers of validity, as outlined above, is of course crucial Gyllstad (2019, p. 388) emphasizes that “validity is not technically a characteristic of a particular test but rather the inferences and interpretations made from the test scores and their use.”

As indicated above the application of test theory is multi-layered and more detailed information on the individual steps taken during test development in this thesis can be found in the chapter 7.

3.2 Existent vocabulary tests

In the following section a selection of the most widely used vocabulary tests measuring the four word knowledge components under investigation is introduced. Moreover, a critical view on some aspects of testing and assessment in the area of vocabulary knowledge is presented and the tests reviewed are related to the measurement instruments used in this study. Finally, while giving an overview of the main vocabulary tests in use it has to be mentioned that the field of vocabulary testing and assessment is flourishing and that more recent tests appear regularly. Therefore, this chapter does not claim to present a complete picture of all tests existent.

Form-meaning knowledge

One of the most widely used tests of form-meaning recognition knowledge is the Vocabulary Levels Test (VLT), which has been originally developed by Nation (1990) to measure receptive vocabulary size. Based on the General Service List (West, 1952) and the University Word List (Xue & Nation, 1984) the test covers knowledge of the most frequent 2000, 3000, 5000 and 10 000 words in the English language (Schmitt, Schmitt & Clapham, 2001). Unfortunately, very limited information on test validation was available for the first version in 1990, with only one note that advises not to use this test on participants with Roman L1s (Schmitt, Nation & Kremmel, 2020). Read (1998) continued validation work on

the first version of the VLT, reporting that it can be considered reliable overall. However, few other researchers continued examining the test but instead it became widely used as assessment tool in schools and in research (Schmitt, Schmitt & Clapham, 2001). This lack of test validation before publication has been recently criticised by Schmitt, Nation and Kremmel (2020) amongst other researchers.

When completing the test participants are supposed to match the target words to their correct definition as visible in example (3) below:

(3) You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

1 concrete	
2 era	_____ circular shape
3 fiber	_____ top of a mountain
4 hip	_____ a long period of time
5 loop	
6 summit	

Overall 18 questions per level are presented. To ensure quick completion of the test and a large item number, definitions are kept short and the words used in the definitions are of a higher frequency band than the words tested (Nation, 1990, p. 264).

Building on the VLT by Nation (1990) Schmitt, Schmitt and Clapham (2001) have conducted an extensive validation study resulting in the development of two new versions of the VLT, which have a similar format as the one outlined above. However, items related to academic word knowledge were sampled from the more recent AWL (Coxhead, 2000) presenting 30 items per level to improve reliability. Moreover, Schmitt, Schmitt and Clapham (2001, p. 79) found that individual items appeared to work well providing a profile which is highly scalable. Moreover, native speakers performed well on the test and personal interviews with the test takers revealed that examinees were taking the test seriously. However, Schmitt, Schmitt and Clapham (2001, p. 79) state that more evidence of validity and reliability of the new versions needs to be collected for different participant groups. This is taken up by Schmitt, Nation and Kremmel (2020, p. 113), who state that more information regarding the use of the new VLT versions for different target groups needs to be provided.

In 2017 Webb, Sasao and Balance continued to further improve the existent versions of the VLT by using the more recent BNC/COCA word list (Nation, 2012) to sample items to enhance the currency of words representing each frequency level. Moreover, they decided to cover five different frequency levels, namely 1000, 2000, 3000, 4000 and 5000 (Webb, Sasao & Balance, 2017, p. 35). To arrive at an estimate of overall receptive vocabulary size they assumed that knowledge of a word entails knowledge of other members of the same word family instead of lemmas, which were used in the prior versions (Webb, Sasao &

Balance, 2017, p. 35). Furthermore, changes were made regarding the presentation of the items. As can be seen in example (4) below a grid provides the test taker with potential items horizontally and vertically supplies the definitions available (Webb, Sasao & Balance, 2017, p. 37). To answer the item the test taker has to check the correct item box for each definition.

(4)

	boy	Rent	report	size	station	thing
how big or small something is						
place buses and trains go to						
young man						

As Schmitt, Schmitt and Clapham (2001) also Webb, Sasao and Balance (2017) conducted extensive test validation, revealing satisfactory results for the participant group tested.

A similarly popular test on receptive form-meaning recognition is the Vocabulary Size Test (VST) (Nation & Beglar, 2007). It provides an estimate of participants' knowledge from the 1st 1000 to the 14th 1000 word families in the English language (Nation & Beglar, 2007, p. 9). Similar as the VLT it uses the BNC word list (Nation, 2006) as source of items. However, the VST has no separate section for academic word knowledge and can be defined as a proficiency test instead of a diagnostic test (Nation & Beglar, 2007, p. 10). Another difference between the tests is the test format. In contrast to the VLT the VST presents 140 items, ten from each frequency band, in a meaning reduced sentence and gives the test taker options to choose from as can be seen in example (5). Therefore, it uses a multiple choice test format in comparison to the multiple matching format of the VLT.

(5) *miniature*: It is a **miniature**.

- a a very small thing of its kind
- b an instrument for looking at very small objects
- c a very small living creature
- d a small line to join letters in handwriting

To answer the test items correctly Nation and Beglar (2007, p. 11) argue that learners "need to have a moderately developed idea of the meaning of the word", making the VST slightly more demanding than the VLT. While the version shown above is monolingual several bilingual versions of the VST have been developed by now (e.g. Elgort, 2013; Nirattisai & Palanukulwong, 2016). However, blind guessing of participants could inflate scores of the VST making it less reliable, which has been considered to be counteracted by including an "I don't know" option (Stoeckel, Bennett & McLean, 2016).

A more recent form-meaning recognition test using a multiple choice format is the VocabLab Test (Peters, Velghe & Van Rompaey, 2019). This test measures learners word knowledge up to the 5000-level of English and French. Therefore, the test is specially geared towards the target group of Dutch-speaking learners in Flanders. While using the multiple choice format found in prior tests the item selection is based on recent frequency lists (Peters, Velghe & Van Rompaey, 2019, p. 61). Adopting Stoeckel, Bennett and McLean's (2016) suggestion to counteract bias through blind guessing, Peters, Velghe and Van Rompaey (2019) have also included an "I don't know option". First evidence for reliability and construct validity were given with test scores decreasing with increasing frequency level. Moreover, the test seems to discriminate well between different proficiency levels (Peters, Velghe & Van Rompaey, 2019, p. 53).

Another widely used format to test receptive vocabulary size is the Yes/No format. In this test format learners are presented with target words for which they need to indicate whether they know the word by clicking yes or whether they don't know the word by clicking no as can be seen in example (6) (Daller, Milton & Treffers-Daller, 2007; Eyckmans et al., 2007, p. 59).

(6)

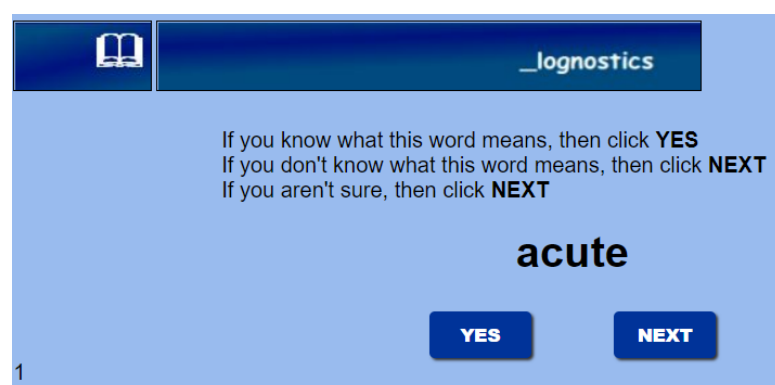


Figure 3.2: Sample item of a Yes/No test format

One of the benefits of this test format is its simplicity and the possibility to test a large number of words in a short time. To account for potential guessing a proportion of pseudowords is mixed with the real English words tested and participants get point reduction for clicking yes for a pseudoword. However, there have been extensive debates about such correction formulas (e.g. Huibregtse, Admiraal & Meara, 2002; Kremmel & Schmitt, 2016; Stubbe, 2012; Stubbe & Steward, 2012). Moreover, difficulties regarding scoring the test as well as the presentation of items out of context were criticized (Meara & Miralpeix, 2017, p. 115). Moreover, it was debated whether Yes/No tests measure the same type or degree of word knowledge as the VLT or the VST, measuring rather knowledge of receptive written word form instead of form-meaning recognition. While learners are indeed tested on whether they can correctly recognize an English written word form, their existent knowledge about the meaning of this word is relying on self-report data and cannot

be proven with this test format. Some tests using this rather widely debated test format are the V_YesNo test (Meara & Jones, 1990) and the X-Lex test (Meara & Milton, 2003).

While all tests for receptive vocabulary size mentioned above are targeted at adults or at least literate participants, the Picture Vocabulary Size Test (PVST) has been primarily developed for use with young pre-literate language learners as a formative, diagnostic test (Anthony & Nation, 2017). The test is based on the most frequent 6000 word families of English for young native speaking children (Anthony & Nation, 2017). In this test format learners can listen to target words and select a matching picture representing the meaning of the word out of four options as visible in example (7).

(7)

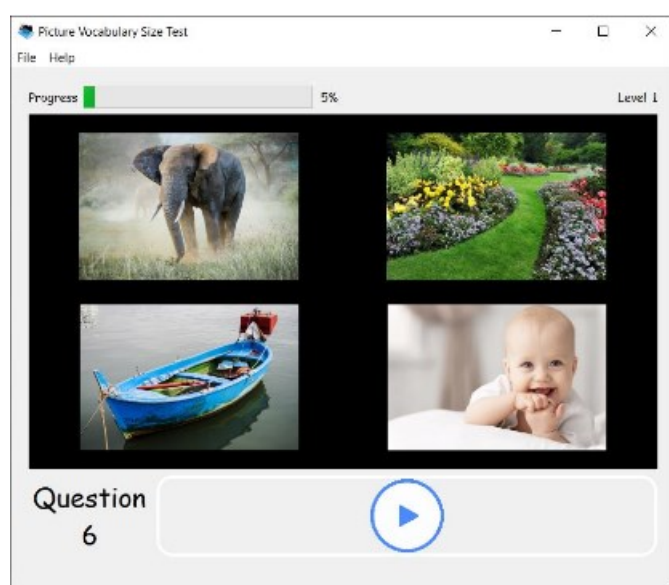


Figure 3.3: Example of an PVST item

The learner can listen to the word twice before choosing a picture or choosing the I'don't know option, if activated (Anthony & Nation, 2017). Hence, while measuring form-meaning recognition knowledge this test presupposes knowledge of a words spoken form in contrast to prior tests, which build on knowledge of the written form. When interpreting the test score one correct answer corresponds to a vocabulary size of 62.5 words (Anthony & Nation, 2017).

Another interesting test format for young learners developed by Sylvén and Sundqvist (2016, p. 5) is the Young Learners Vocabulary Assessment Test (YLVAT), which covers 37 items ranging between K1 and K2 level. The test consists of three different parts to counteract test fatigue. The first part testing word recognition comprises 13 items integrated into statements, which have to be answered with true, false or not given. The second part consists of twelve items taken from the VLT, which should test form-meaning recognition knowledge. Lastly, the third part leads over to form-meaning recall knowledge, since it covers twelve items of the Productive Vocabulary Levels Test (Sylvén & Sundqvist, 2016, p.

5), which will be introduced at a later point. As their validation study covering 52 L1 Swedish speaking secondary school students aged twelve years shows, the test discriminates well between the different levels of ability existent in this participant group (Sylvén & Sundqvist, 2016, p. 8). Moreover, a combined elevation of the scores on the national test of English reveal further evidence for test validity (Sylvén & Sundqvist, 2016, p. 12).

Similar to the YLVAT the Computer Adaptive Test of Size and Strength (CATSS) is combining receptive and productive measurements of vocabulary knowledge (Aviad-Levitzky, Laufer & Goldstein, 2019). However, the test follows a slightly different conceptualization by not only differentiating between recognition and recall but additionally between the recognition or recall of form or meaning of a word. When taking a look at example (8) provided below reveals that the test comprises four different parts measuring either recall of word form, recall of word meaning, recognition of word form or recognition of word meaning (Aviad-Levitzky, Laufer & Goldstein, 2019, p. 346).

(8) Recall of word form: She is a / _____ girl. (small)

Recall of word meaning: 'She is a **little** girl' means that she is _____

Recognition of word form: She is a _____ girl. (small)

a. little b. great c. nice d. single

Recognition of word meaning: 'She is a **little** girl' means that she is _____

a. small b. great c. nice d. single

While using fill-in-the-gap formats to measure both aspects of word recall, once providing the word and once a meaning expressed through a synonym or definition, a single-choice format is applied for the recognition parts (Aviad-Levitzky, Laufer & Goldstein, 2019). Based on Laufer et al.'s (2004) and Laufer and Goldstein's (2004) research the CATSS test assumes an implicational scale starting with the most difficult level being recall of word form over recall of meaning to recognition of form and recognition of meaning. Consequently, if a word is already known in the recall of word form test section, the word is not tested in the other modalities anymore (Aviad-Levitzky, Laufer & Goldstein, 2019, p. 346). If no or an incorrect answer is given the word is tested again in the next lower modality. Therefore, not all words are tested in all four modalities saving proficient test takers' time.

Regarding sampling of words the CATSS test covers 14 frequency bands with ten words per frequency band selected to represent the level. While Aviad-Levitzky, Laufer and Goldstein (2019, p. 349) have based this decision on Beglar's (2010) result that ten words are sufficient to present a frequency band, recent findings suggest that at least 30 words should be included (Gyllstad et al., 2015; Schmitt, Schmitt & Clapham, 2001). Moreover, it has to be clarified that the CATSS test presupposes knowledge of grammatical accuracy and orthography of words tested (Aviad-Levitzky, Laufer & Goldstein, 2019, p. 348). However, this decision already implies an acquisition order of grammatical accuracy and orthography being learned before recall of word form; or at least that recall of word form is only known,

if orthography and grammatical accuracy are met. This presupposition appears to be difficult when considering the limited research existent regarding the acquisition order and relationship between different word knowledge components. Furthermore, this might enhance the difficulty level of the test drastically for particular groups of test takers.

One of the models for the CATSS test (Aviad-Levitzky, Laufer & Goldstein, 2019) is the Productive Vocabulary Levels Test (PVLТ) which has been developed as a complementary part to the VLT (Nation, 1990; Schmitt, Schmitt & Clapham, 2001; Webb, Sasao & Balance, 2017) measuring form-meaning recall instead of recognition knowledge (Laufer & Nation, 1999; Nizonkiza & van den Berg, 2014). Similar to the VLT (Nation, 1990; Schmitt, Schmitt & Clapham, 2001; Webb, Sasao & Balance, 2017) it samples 18 items at the 2000, 3000, 5000, 10 000 level and from the University Word List (Xue & Nation, 1984). The test makes use of a fill-in-the-gap format as visible in example (9) below.

(9) The book covers a series of isolated **epis** ____ from history.

In the test each item is presented in a meaningful sentence context with the first letters of the word given to restrict the options suitable for the gap (Laufer & Nation, 1999). Similar to the VLT (Nation, 1990; Schmitt, Schmitt & Clapham, 2001; Webb, Sasao & Balance, 2017) the PVLТ is fast and easy to administer and allows for testing a large number of words (Nizonkiza & van den Berg, 2014). However, whereas it was shown that the test could distinguish between learners with different levels of language proficiency (Schmitt, Nation & Kremmel, 2020, p. 113), some potential flaws of the format can be identified. As Meara and Fitzpatrick (2000, p. 21) highlight the contextual information in which the target item is presented is providing the reader with varying degrees of information making some items harder to take than others. Moreover, Read (2000) has noted that reading skills might potentially interfere with the test scores, since test performance is relying on the learners to understand the surrounding sentence context. A very similar problem is caused by the fact that the PVLТ has to present a varying number of letters given for individual words to counteract ambiguity of answers, which again influences the difficulty of items (Read, 2000, p. 125). This problematic aspect will be further addressed in chapter 7.1 since a similar format has been used for the Form-meaning recall test of this study.

Overall, all vocabulary tests reviewed so far try to give an estimate of learners' overall receptive or productive vocabulary size by multiplying the scores on the test by either the number of lemmas existent for the word or the number of word family members for the base word tested. However, research implies that the counting unit of a word family is a poor indicator of knowledge of all derivational forms (Browne, 2013; McLean, 2018; Stoeckel, McLean & Nation, 2021). Consequently, vocabulary size tests which assume that knowledge of the base form of a word family infers knowledge of all inflectional and derivational forms of this word tend to overestimate learners' vocabulary size (Stoeckel, McLean & Nation, 2021). In comparison tests basing their counts on the knowledge of

lemmas appear to have less but still existent bias regarding the overall score (Stoeckel et al., 2018).

Another issues worth addressing is that all vocabulary size tests additionally work with the assumption that choosing a small sample of words per frequency band and then multiying these will give a sufficient indication of learners' knowledge of this frequency level and consequently their overall vocabulary size. This again is based on the idea that words in the same frequency band are of similar difficulty. However, research has shown that frequency only predicts which words are known to a limited extent and that other factors additionally influence a word's perceived difficulty and when it is going to be learned (Hashimoto, 2021; Schmitt, Nation & Kremmel, 2020; Stoeckel, McLean & Nation, 2021).

Based on these issues this research study refrained from developing and using a classical vocabulary size test. While using similar design formats for the form-meaning recognition and the form-meaning recall test, no estimates of vocabulary knowledge beyond the words tested have been made. Instead form-meaning recognition and recall knowledge are not primarily seen as a representation of vocabulary size but are viewed as one equal aspect of word knowledge amongst other word knowledge components. Therefore, the purpose of the tests developed for this study is rather to compare knowledge of different word knowledge components of the same words.

Multiple meaning knowledge

Concerning multiple meaning recognition knowledge a smaller sample of tests is available for different educational levels, which have not been implemented in a standardized manner.

The Receptive Polyseme Vocabulary Test (Booton et al., 2021) is a test to measure knowledge of multiple meanings of children in early primary school years by using a similar format than the PVST (Anthony & Nation, 2017). As visible in the example (10) below the test takers hear a target word, such as *train* or *groom* during the test and need to select two out of six pictures which represent two different meanings of the word heard (Booton et al., 2021, p. 171).

(10)

Word	Targets		Distractors			
	Meaning A	Meaning B	Meaning A associate	Meaning B associate	Phonological associate	Other word
train					 "drain"	
groom					 "game"	

Figure 3.4: Example of an item of the Receptive polyseme vocabulary test

While one of the distractors is semantically completely unrelated to the target meanings, two distractors are associates of the two meanings in question and one is a phonologically similar word (Booton et al., 2021, p. 172). Overall the test covers 30 homonyms of medium to high frequency and 1-2 syllable length. The test was assessed for reliability (test-retest & internal consistency) as well as convergent validity with other vocabulary tests by the authors (Booton et al., 2021, p. 175).

Moving on to assessing multiple meaning recognition knowledge at secondary school level the Polysemous Word Test has been developed to measure bilingual adolescents' knowledge of polysemous vocabulary (Logan & Kieffer, 2017, p. 1688) and how this knowledge might be related to reading proficiency. The test covers two meanings of 21 words, which are presented in two different sets of sentences as visible in example (11) below.

(11) I scratched my head in confusion.	The general is the head of the army.
In this case, does <i>head</i> mean:	Does <i>head</i> mean:
a. cabbage	a. cabbage
b. sneaker	b. sneaker
c. part of the body	c. part of the body
d. leader of a group	d. leader of a group
e. to go somewhere	e. to go somewhere

For each word a casual meaning, representing the every day use of the word and an academic use of the word are presented (Logan & Kieffer, 2017, p. 1693). To sample items with two different meanings WORDNET, a linguistic database of the university of Princeton was used (Logan & Kieffer, 2017, p. 1693). In a next step casualness and imagiability ratings for the word meanings from the MRC Psycholinguistic Database as well as likelihood of use in academic register were used to identify the casual and the academic meaning of the words (Logan & Kieffer, 2017, p. 1694). Overall, word meanings with higher imagiability and lower academic register scores were considered to be the casual meaning of a word (Logan & Kieffer, 2017, p. 1694).

For each of the meanings test takers have to identify the target words meaning in the sentence out of five possible choices (Logan & Kieffer, 2017, p. 1694). While one out of the five options is correct, at least one distractor is an alternate but inappropriate meaning of the target word, at least one is an associate but not synonymous term and at least one is unrelated to the target word but is meaningful in the sentence context (Logan & Kieffer, 2017, p. 1694). The authors have tested for internal reliability and convergent validity with other vocabulary tests and language proficiency measures in their data set. However, they

recommend to double check the tests reliability and validity in other contexts (Logan & Kieffer, 2017, p. 1695).

At the tertiary level Qian's (2002, p. 520) Depth of Vocabulary Knowledge Measure (DVK) was developed to assess students' knowledge of word meanings and the link to reading comprehension. The DVK is heavily based on the Word Associates Test by Read (1993, 1995), which has also been used as measure for collocation recognition knowledge and will be described in more detail in the next section. Using 32 of the original items of the Word Associates Test (Read, 1993, 1995) and adding eight new items, the DVK (Qian, 2002, p. 524) covers 40 items in total. The target words chosen are "general academic adjectives" (Qian, 2002, p. 525). As visible in example (12) below each item consists of one target word presented above two boxes. While the left box is offering words which might be synonymous to one meaning aspect or the whole meaning of the stimulus word, the right box offers potential collocates of the word (Qian, 2002, p. 524).

(12)

<i>sound</i>	
logical healthy bold solid	snow temperature sleep dance

The task of the test taker is to identify one to three correct meaning related words from the left box and one to three correct collocates from the right box. Overall, four answers per item are correct with different distribution patterns of correct answers possible (Qian, 2002, p. 524). Regarding scoring each correct answer receives one point, while incorrect answers are not penalised. Therefore, per item four points are possible, summing up to 160 points for the whole test (Qian, 2002, p. 525). Similar to prior tests internal reliability and convergent validity with other vocabulary and language proficiency tests were checked by the author with Korean and Chinese speakers participating in the study (Qian, 2002, p. 528).

Another test measuring multiple meaning knowledge of tertiary level students is the Vocabulary Depth Test (Richard, 2011). However, as visible in example (13) below this test can be considered to be a recall test of multiple meaning knowledge in comparison to recognition tests mentioned before (Richard, 2011, p. 107). As the example shows participants are presented with six sample sentences of the same target word, which has to be correctly inserted into the gap (Richard, 2011, p. 111).

(13)

- i. It's _____ a small cut.
- ii. My house is _____ around the corner.
- iii. It's _____ the right thing to do.
- iv. We _____ arrived.
- v. _____ do it.
- vi. _____ the two of us.

All 50 items completed by the participants in Richard's (2011, p. 110) study were chosen by random sampling from the 1K list. In a next step six sample sentences were created for each of the target words using example sentences given in the Collins Cobuilds Advanced Learners dictionary (Sinclair, 2003). For correct answers points were awarded while no points were deducted for incorrect answers (Richard, 2011, p. 111). Moreover, in contrast to other productive measures this test accepted spelling errors. Similar to prior tests Richard (2011, p. 107) evaluated the test's reliability and convergent validity by comparing the score to the performance on the VKS (Nation & Beglar, 2007). Moreover, two proficiency groups were compared and the test seemed to distinguish well between the different proficiency levels (Richard, 2011, p. 107).

Lastly, another test format used by González-Fernández and Schmitt (2019, p. 5) to measure multiple meaning recognition knowledge of 144 Spanish learners of English at different proficiency levels will be presented due to a similar format used in this study. Since this test has been part of an extensive test battery investigating several different word knowledge aspects it only covers 20 target words from a range of frequency levels (1K-9K) (González-Fernández & Schmitt, 2019, p. 5). As visible in example (14) below the format adapted from Li and Kirby (2015) presents participants with five different sentences, three of them using the target word with existent meanings and two being distractors (González-Fernández & Schmitt, 2019, p. 9).

(14)

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

Therefore, test takers need to correctly identify the existent meanings of the target word, which could be both homonymous or polysemous senses (González-Fernández & Schmitt, 2019, p. 6). While three different meanings of a word were a prerequisite for target words participants were informed that one to three of the sentences could be correct to minimize guessing (González-Fernández & Schmitt, 2019, p. 9). Unfortunately, no information about piloting results or test validation of the test battery in use could be found.

Collocation knowledge

When it comes to testing knowledge of multiword units time has shown that testing formulaic sequences appears to pose more of a challenge than testing single words (Gyllstad & Schmitt, 2018). Due to the complexity of the construct of multiword units or the manifold definitions in the field, amongst other reasons, test development remains to be difficult (Gyllstad & Schmitt, 2018). Hence, no real standardized vocabulary tests for

multiword units have been established yet. However, the following tests can be repeatedly found in the literature.

The Word Associate Tests (WAT) (Read, 1993; Read, 1998) is one of the most established tests covering multiword units in the field. Following the format of a multiple choice test the WAT states to investigate both knowledge of word meaning, in the form of synonym knowledge and collocations. Covering 40 items, the test presents the test takers with four potential synonyms on the one side and four potential collocations on the other side from which the test takers should choose the correct options (see example 15 below).

(15)

1. beautiful

☐ enjoyable	☐ expensive	☐ free	☐ loud	☐ education	☐ face	☐ music	☐ weather

Using an item response theory (IRT) Rasch analysis, Read (1993) reports good reliability for this test with a Cronbach α of around 0.9. However, validation shows that consistent and suitable scoring of the test poses particular challenges.

Gyllstad's (2007, 2009) COLLEX format similarly presents its 50 items in a decontextualized manner by asking the participants to choose one out of three options which is most conventional and frequent in the English language as visible in example (16) below.

(16)

One of the three word combinations in each item is a natural and frequent word combination occurring in the English language, whereas the other two are not. Choose the word combination you think is the most natural and frequently occurring by ticking the box that corresponds to it in the right margin.

1	a. do damage	b. make damage	c. run damage	<table border="1"> <tr> <td>a</td> <td>b</td> <td>c</td> </tr> <tr> <td>☐</td> <td>☐</td> <td>☐</td> </tr> </table>	a	b	c	☐	☐	☐
a	b	c								
☐	☐	☐								

However, while the WAT covers adjective + noun collocations the COLLEX investigates verb + noun phrase collocations, which were selected based on corpus analysis and the z-score, a measure of collocational strength already introduced in chapter 2.2.1. Similar to the WAT the COLLEX reports a high reliability with a Cronbach α of 0.89 and has been administered with 269 participants.

Both test formats, the WAT (Read, 1993; Read, 1998) and the COLLEX (Gyllstad, 2007), have been used as models for the two versions of the Collocation Recognition Test of Academic Verbs developed for this study. Therefore, the formats will be addressed again with an eye on the methodology in chapter 7.3.

A very similar format is the Discriminating Collocation Test (DISCO) (Eyckmans, 2009) which also covers 50 items but asks participants to select two out of three idiomatic verb+noun combinations as visible in example (17).

(17)

Tick the two phrases that are the correct idioms in English:

- seek advice
- express charges
- pay attention

Similar to the COLLEX it covers high-frequency verbs, which have been selected on the basis of corpus analysis and the z-score. However, Eyckmans (2009) divides her target words into three frequency categories to investigate the effect of frequency on multiword level. Corresponding to the tests listed before a high reliability of 0.9 is reported additionally to the finding that frequency of collocation was a significant predictor of collocation knowledge.

A related test also developed by Gyllstad (2007; 2009) is the COLLMATCH, which can be described as a yes/no test, asking participants to indicate whether they believe a word combination is existent in the English language or not as visible in example (18) below.

(18)

1	have a say	2	lose sleep	3	do justice	4	draw a breath
☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej

As for other yes/no test formats the test contains pseudo collocations, which can be used to adjust for the effect of guessing. The ratio of the COLLMATCH is 30 pseudowords out of 100 items. Congruent to the COLLEX the COLLMATCH format has high reliability of $\alpha = 0.89$. Moreover, construct and concurrent validity are looked at by examining the correlation between COLLEX and COLLMATCH ($r = 0.86$, $p < 0.01$), which shows that the two tests are significantly and clearly related. The test format of the COLLMATCH can be said to be widely related to the older Association Vocabulary Test (Vives Boix, 1995). However, this test claims to tap into both constructs of vocabulary size and depth by presenting participants with word pairs and asking them to decide whether they are associated or not (Vives Boix, 1995).

Another interesting format testing collocational knowledge is the CONTRIX (Revier, 2009; 2014), which also covers 45 verb + noun combinations. However, in contrast to the decontextualized tests before Revier (2009; 2014) presents his items in the context of a short sentence with a gap as visible in example (19) below, which needs to be filled with the correct verb+ noun collocation.

(19)

Choose one word from each column to make an appropriate phrase to complete the sentence.

The quickest way to win a friend's trust is to show that you are able to

tell	a/an	joke
take	the	secret
keep	---	truth

To fill the gap the participants can choose a verb + a determiner if necessary + a noun from a three column table. All collocations covered are either non-, semi- or transparent collocations of 15 high-frequency verbs. Similar to the other tests before Revier (2009; 2014) reports good reliability.

Another test of multiword units which is in line with the CONTRIX (Revier, 2009; 2014) is the Phrasal Vocabulary Size Test (PHRASE) (Martínez, 2011). This test covers 505 phrasal expressions selected from the PHRASE list (Martínez & Schmitt, 2012) and presents them in a non-defining sentence context comparable to the CONTRIX (Revier, 2009, 2014), asking participants to select from multiple fill-in options as visible in example (20).

(20)

First 1000 [\[Go 2\]](#)

1. GO ON: It will **go on**.

- a. ☐ sleep
- b. ☐ repeat
- c. ☐ be fast
- d. ☐ continue

However, since the target words have been selected from an existent list testees' results can be expressed in percentages in comparison to the results of other tests (Gyllstad & Schmitt, 2018).

Moreover, several more collocation tests have been developed, such as the V_Links (Meara, 2009), the phrasal vocabulary size test (Martínez, 2011) or the receptive knowledge of collocations test (Nguyen & Webb, 2017). In addition, new measurement approaches to collocation knowledge are taken by more psycholinguistic tests, which use concepts, such as processing time as indicators for participants knowledge (Gyllstad, 2019).

4 Literature review vocabulary knowledge

While the prior chapter has outlined quality criteria for tests and has introduced existent vocabulary tests for the four word knowledge components investigated in this study, this chapter will present an overview of prior studies concerned with these word knowledge

aspects. Since research on vocabulary knowledge is expanding dramatically this literature review, while giving a good overview of the field and allowing comparison of findings, does make no claims regarding completeness. Table 4.1 below gives an overview of the studies considered for the following literature review.

	Form-meaning recognition	Form-meaning recall	Multiple meaning recognition	Collocation recognition
2022	-	- Ghamarian	- Söderström, Thorén & Karlsson	-
2021	- Logan & Kieffer - Szabo, Stickler & Adinolfi	- Li & Hafner - Asaad & Shabdin	- Booton et al. - Logan & Kieffer - González- Fernández	-
2020	- Duff & Brydon - Lee - Schwarz - Skjelde & Coxhead	-	- Alqraini & Paul - Sonbul, El-Dakhs & Al-Otaibi	- Shabani & Rahimy - Kamarudin, Abdullah & Abdul Aziz - Vafae & Suzuki
2019	- Lawrence et al. - Yang et al. - González- Fernández & Schmitt - Dabbagh & Enayat - Peters, Velghe & Van Rompaey	- Yang et al. - Uchihara & Saito - Lee (a) - Lee (b)	- Zhang & Wen - Abou-Khalil et al. - González- Fernández & Schmitt	- Dentisak - Dabbagh & Enayat - Vilkaité & Schmitt
2018	- Cheng & Matthews - Sun et al. - Miralpeix & Muñoz - Coxhead & Boutorwick - Mondria & Wiersma - Masrai & Milton	- Cheng & Matthews - Zhong - Sun et al. - Miralpeix & Muñoz - Mondria & Wiersma - McKeown et al.	- Logan	
2017	- Binder et al. - Hahn - Zichtl	- Fitzpatrick & Clenton - Zichtl	- Alnamer - Logan & Kieffer	- Talakoob & Koosha - Nguyen & Webb

	- Luxton, Fry & Coxhead - Makhoul - Roghani & Milton	- Schuth, Köhne & Weinert - Makhoul - Tahmasbi & Farvardin		- Özdem Ertürk
2016	- Townsend et al. - Sylvén & Sundqvist - Ozturk	- Sylvén & Sundqvist - Nizonkiza	- Garnier & Schmitt	- Szudarski & Carter
2015	- Zhang & Lu - Li & Kirby - Ma & Lin - Coxhead, Nation & Sim - Ozturk	- Olsson & Sylvén - Ozturk	- Li & Kirby	- Miščin - Adelian, Nemati & Fumani - Nizonkiza - Ma & Lin
2014	- Pham & Khonert - Wang - Yamamoto - Sylvén & Ohlander - Peters	- Yamamoto - Pham & Khonert	-	- Noori, Gholami & Rajabi - Begagić - Wang
2013	- Roche & Harrington	-	- Karlsson	- Wolter & Gyllstad - Gyllstad - Webb, Newton & Chang - Sonbul & Schmitt
2012	- Townsend et al. - Webb & Chang - Elgort - Ehsanzahdeh	-	-	-
2011	-	-	-	- Alsakran - Richard
2010	- Marchman, Fernald & Hurtado	-	-	-
2009	- Webb - Llach & Gallego - Sundqvist - Staehr - Orosz - Milton	- Vagh, Pan & Mancilla-Martinez	-	- Staehr - Webb & Kagimoto - Li & Schmitt

2008	- Webb - Staehr	-	-	-
2007	- Miralpeix	-	- Morimoto & Loewen	- Gyllstad
2006	- Jiménez Catalán et al.	-	- Sullivan	- Barfield
2005	-	-	-	-
2004	- Gyllstad - Tschirner	-	-	-
2003	-	-	- Verspoor & Lowie	-
2002	- Ordóñez et al. - Cameron - Qian	-	-	- Qian
2001	-	-	-	-
2000	- Fan - Cobb	- Fan	-	-
1999	- Ishihara, Okada & Matsui - Nurweni & Read	- Ishihara, Okada & Matsui	-	- Nurweni & Read
1998	- Laufer	- Laufer	- Schmitt	-
1997	- Schmitt & Meara	-	-	-
1996	-	-	-	- Bonk
Total number	63	28	19	31

Table 4.1: Summary of studies considered in the literature review on vocabulary knowledge

4.1 What we already know about form-meaning recognition knowledge

As visible in Figure 4.1 below form-meaning recognition knowledge has been widely researched in the field.

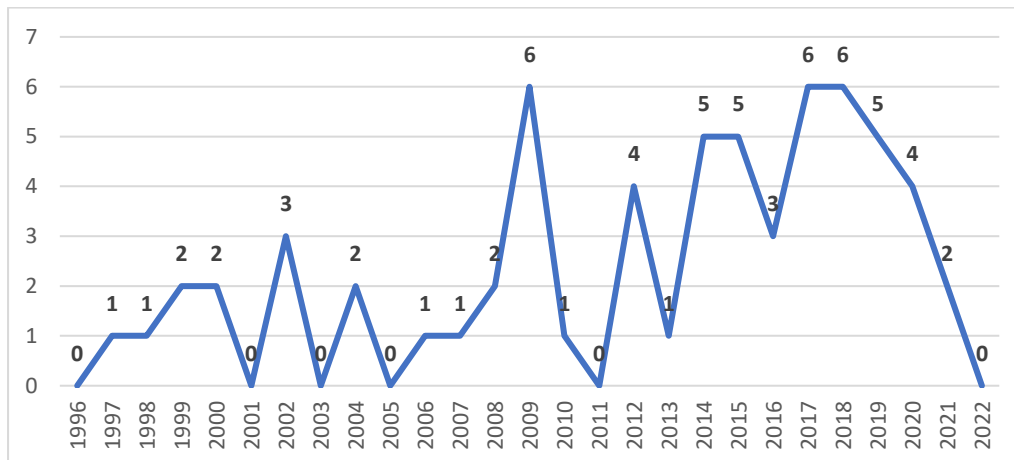


Figure 4.1: Distribution of publication year of form-meaning recognition studies

The majority of the studies considered were published after 2009, which highlights the aim of presenting the most recent findings of the field.

As presented in chapter 3.2 above there is a relatively large variety of available form-meaning recognition and written form recognition tests, which are the two most common word knowledge components measured to estimate receptive vocabulary size. This is also reflected in the type of tests applied in the studies reviewed as depicted in Figure 4.2 below.

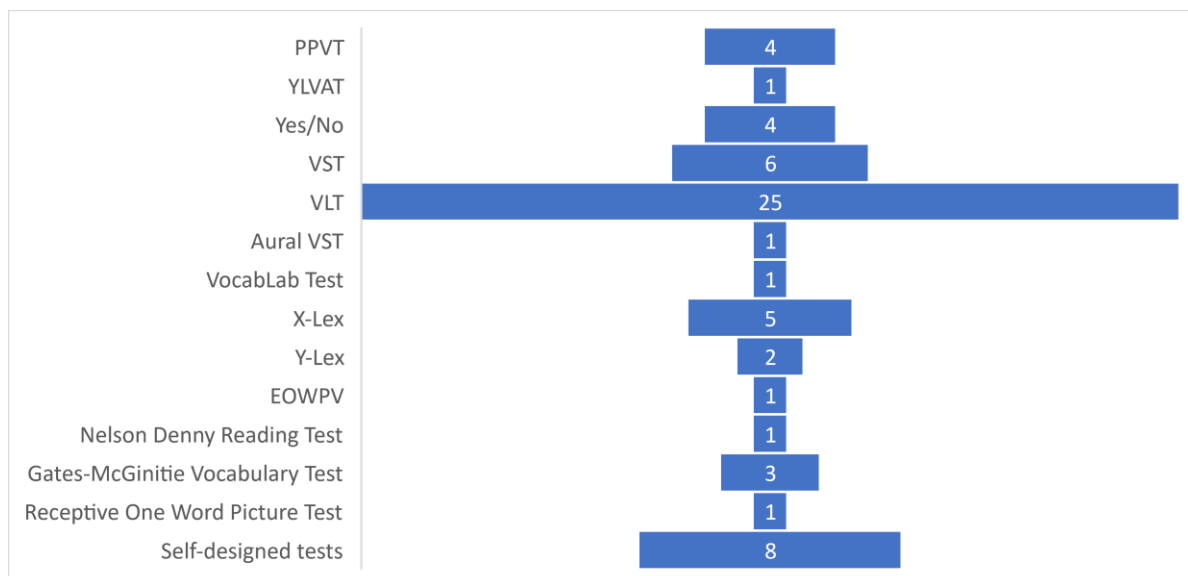


Figure 4.2: Distribution of testing method in receptive vocabulary size studies reviewed

As visible, most studies interested in receptive vocabulary size make use of the VLT (Nation, 1990; Schmitt, Schmitt & Clapham, 2001), which is a form-meaning recognition test with a similar format to one of the tests used in this study. In total, the VLT was utilized 25 times in

the studies reviewed. Moreover, the VST (Nation & Beglar, 2007) was used as measurement instrument six times, making it the second most prominent test in the reviewed literature. This is followed by the X-Lex (Meara & Milton, 2003), which was used five times to estimate receptive vocabulary size. However, it has to be mentioned that the X-Lex, using a yes/no test format, is rather a measure of written form recognition as already outlined in chapter 3.2, which limits the comparability of results to studies using form-meaning recognition tests. The same accounts for the three occasions when a Yes/No test (Meara, 2010) was used for measurements. Similar, the PPVT (Dunn & Dunn, 2007), which was used at least four times, has to be compared to other studies with caution, since the PPVT traditionally is an aural form-meaning recognition test. This requires the participants to link aural clues to (one of) their potential meaning/s. Moreover, vocabulary sections of commercially available tests, such as the Nelson Denny Reading test (Brown, Fishco & Hanna, 1993) and the Gates-McGinitie Reading test (https://riversideinsights.com/gates_macginitie) were used. Apart from widely used tests a surprisingly large number of authors decided to develop tests specifically for their studies. This has the advantage that the tests are tailor-made for the intended audience. Nonetheless, tests are not always checked for their reliability or validity (Kremmel & Schmitt, 2016, p. 377). Hence, this study, which also bases its results on tailor-made tests, pays specific attention to consistent development criteria and validation evidence.

Regarding features of the participant groups researched it can be stated that most studies concerned involve participants from either America, Asian countries or European countries as visualized in Figure 4.3 below.

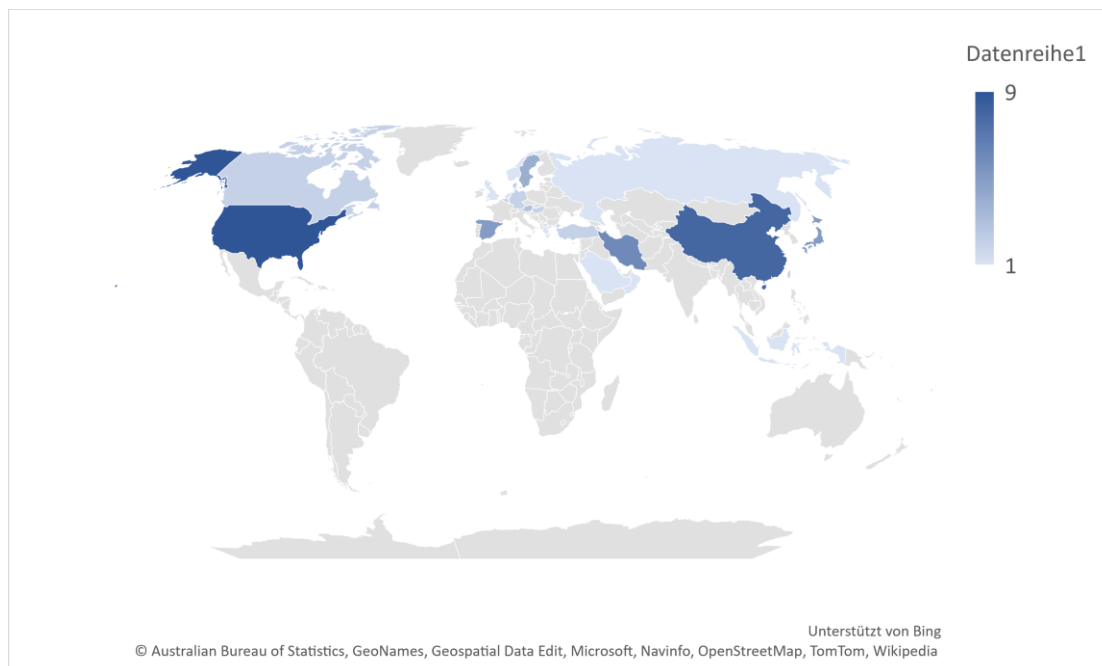


Figure 4.3: Geographical spread of form-meaning recognition studies reviewed

As visible most studies on form-meaning recognition were conducted in the United States of America, with nine studies involving bilingual learners or EFL learners learning English in a native speaking environment (e.g. Binder et al., 2017; Duff & Brydon, 2020; Lawrence et al., 2019; Logan & Kieffer, 2021; Marchman, Fernald & Hurtado, 2010; Ordóñez et al., 2002; Pham & Khonert, 2014; Townsend et al., 2016; Townsend et al., 2012). Moreover, Asian countries are very prominent, with eight studies involving participants from different regions of China (e.g. Cheng & Matthews, 2018; Fan, 2000; Li & Kirby, 2015; Sun et al., 2018; Wang, 2014; Yang et al., 2019; Zang & Lu, 2015), five studies examining form-meaning recognition knowledge of Japanese participants (Ishihara, Okada & Matsui, 1999; Schmitt & Meara, 1997; Webb, 2009; Webb, 2008; Yamamoto, 2014), two studies with Taiwanese participants (Ma & Lin, 2015; Webb & Chang, 2012) and one study each located in Indonesia (Nurweni & Read, 1999) and Korea (Lee, 2020). In comparison European countries appear to be similarly represented with five studies in Spain (Jiménez Catalán et al., 2006; Llach & Gallego, 2009; Miralpeix, 2007; Miralpeix & Muñoz, 2018), four studies in Sweden (Gyllstad, 2004; Sundqvist, 2009; Sylvén & Ohlander, 2014; Sylvén & Sundqvist, 2016) three studies in Austria (Hahn, 2017; Schwarz, 2020; Zichtl, 2017), two studies in Belgium (Peters, 2014; Peters, Velghe & Van Rompaey, 2019), two studies in Germany (Coxhead & Boutorwick, 2018; Tschirner, 2004), two studies in Denmark (Staehr, 2009; Staehr 2008), one study in Norway (Skjelde & Coxhead, 2020), one study in Hungary (Orosz, 2009), one study in Great Britain (Cameron, 2002), one study in Greece (Milton, 2009) and one study in the Netherlands (Mondria & Wiersma, 2018). This amounts to 23 studies based in a European context and 17 studies with an Asian background. Hence, the sample of studies is rather balanced geographically, overall including participants from 24 different countries. Only African countries appear to be strongly neglected unfortunately concerning research on form-meaning recognition knowledge. A more detailed map can be found in appendix 4.1 together with maps for the other three word knowledge components researched.

Concerning the educational level of the participants researched, most studies examined form-meaning recognition knowledge of participants completing their tertiary education, with 30 studies researching form-meaning recognition knowledge of university students as visible in Figure 4.4 below.

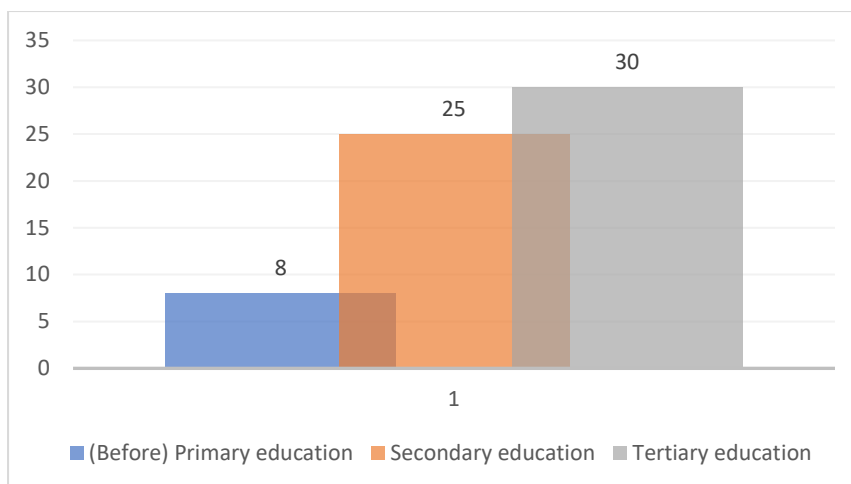


Figure 4.4: Educational level of participants involved in form-meaning recognition studies reviewed

Yet, studies on participants involved in tertiary education are closely followed by 25 studies examining form-meaning recognition knowledge of secondary school students. Only primary education appears to have been less concerned, which points to an interesting research gap to be filled by future studies, considering the increasing influence of EE activities also at primary school age. In the following research studies concerning either primary, secondary or tertiary level education are going to be reviewed separately, considering the developmental nature of vocabulary knowledge and the educational system.

4.1.1 Primary education

While research on vocabulary knowledge at primary school level appears to have been rather neglected in the past, most studies conducted seem to have reached a consensus regarding the development of form-meaning recognition knowledge. Several studies (Milton, 2009; Orosz, 2009, p. 188-189; Pham & Khonert, 2014, p. 776; Sun et al. 2018, p. 16) report a significant growth of form-meaning recognition knowledge during primary school education. While Milton's (2009, p. 79) examination of 227 EFL learners at a primary school in Greece resulted in a form-meaning recognition knowledge growth of 500 words per year, Orosz (2009, p. 187) found a smaller vocabulary growth within 253 Hungarian primary school children, ranging between 280 and 481 words a year. However, these results need to be compared with caution, since they differ greatly in their methodological approach taken. While studies such as Milton (2009) and Pham and Khonert (2014) accompanied the same student population for a certain amount of time, asking them repeatedly to fill in the form-meaning recognition tests, Orosz (2009) decided on a pseudo-longitudinal approach. This means that only one data collection was conducted involving children at different stages of their primary school education, such as class A compared to class B, and comparing the ability of these different groups to each other. Hence, differences and developments detected for these participants might also be attributed, at least partially, to differing proficiency levels of the different groups compared.

Concerning an exact score or form-meaning recognition knowledge estimate, results appear to be more varied. Llach and Gallego (2009, p. 138) discovered an average receptive vocabulary size of 1106 words for their Spanish primary school population of 274 pupils, making use of parts of the VLT, covering only knowledge of the most frequent 2000 words of the English language. In contrast, Jiménez Catalán et al. (2006, p. 26), using the same vocabulary test, reported a slightly smaller receptive vocabulary size of below 1000 words for his participants, being 130 Spanish primary school children. However, these discrepancies might be partially explained by differences in age and educational stage of the participants involved in the studies, as is shown by Orosz (2009, p. 188) who reports knowledge of 348 words in the first year, 696 words in the second year, 1177 words in the third and 1457 words in the fourth year of primary school education. Hence, students at primary school level demonstrate a rather rapid development of their form-meaning recognition knowledge, explaining for divergent findings regarding receptive vocabulary size of a rather large category of primary school pupils. This is also supported by the finding of Sun et al. (2018, p. 17), who state that an age effect, with older children learning more quickly, could be detected for the development of form-meaning recognition knowledge. Therefore, vocabulary estimates might vary dependent on the age and developmental stage of the participants involved in a study. This highlights the importance of stating the exact age of primary school children since dependent on the context primary school education ranges from four to seven years.

4.1.2 Secondary education

Comparing several studies at secondary school level reveals that receptive vocabulary size estimates appear to be comparable across different studies (e.g. Logan & Kieffer, 2021; Luxton, Fry & Coxhead, 2017; Mondria & Wiersma, 2018). Schwarz (2020, p. 233) investigating 201 secondary school students in grade 10, aged between 15 and 16, using the Yes/No Test (Meara, 2015), reports that her participants had a mean receptive vocabulary size of 4830.47 words. While this is the highest estimate of receptive word knowledge in the studies compared, Hahn (2017, p. 81) found a similar estimate within her 83 Austrian 10th grade students, who on average were able to recognize the written form of 3041.43 words. Moreover, an older study by Laufer (1998, p. 262) involving secondary school students in grade 10 and 11, aged between 16 and 17, found a receptive vocabulary size between 1900 and 3500 words known. Hence, the maximum score discovered by Laufer (1998) is comparable to the mean scores reported by Schwarz (2020) and Hahn (2017). However, these discrepancies might be partially explained by the testing methods used. While both Schwarz (2020) and Hahn (2017) based their calculations on the Yes/No test by Meara (2015), which requires participants to identify known written word forms, Laufer (1998) utilized the VLT, which measures form-meaning link recognition. Consequently, this difference in word knowledge aspect measured, although being subsumed under the term

receptive vocabulary size, might make a difference regarding resulting vocabulary size estimates.

Furthermore, one needs to differentiate between vocabulary size studies using tests with different counting units. While the studies mentioned above result in total vocabulary size counts in number of words known, other tools result in an estimate of word families known (Duff & Brydon, 2020; Elgort, 2012; Zichtl, 2017). Here the maximum receptive vocabulary size was achieved by Duff and Brydon's (2020, p. 454) 41 participants, who reached a receptive vocabulary size of 7300-9100 word families. However, this exceptionally high result might partially be attributed to the fact that Duff and Brydon's (2020, p. 464) study is based in the US and includes highly proficient speakers of the English language. This is also reflected by the fact that studies by Zichtl (2017) and Elgort (2012) conducted in EFL countries resulted in lower receptive vocabulary size estimates. Zichtl (2017, p. 65) comparing secondary school students in grade 8, 10 and 12 found a receptive vocabulary size ranging from 6640 word families known in the 8th grade to 8850 word families known in the 12th grade of secondary school. Therefore, Zichtl's (2017) participants being at the end of secondary school education appear to approach the results of Duff and Brydon's (2020) participants in the US. These findings are not fully supported by Elgort (2012, p. 259-264) examining 151 students at the end of secondary school, who only reached a receptive vocabulary size of 5922-6594 word families known.

Additional to estimates made regarding the number of words or word families known, several studies present their results in percentages of correct answers on a test taken (Coxhead, Nation & Sim, 2015; Li & Kirby, 2015; Ordóñez et al., 2002; Peters, Velghe & Van Rompaey, 2019; Staehr, 2008). Similar to total vocabulary estimates of secondary school pupils described above these studies report rather comparable results. Sylvén and Ohlander (2014) investigated the form-meaning recognition knowledge of 119 CLIL and 76 non-CLIL students using the VLT. They found that male CLIL students achieved the highest score with 74.67% followed by female CLIL students and male non-CLIL students with 71.3% correct answers and lastly non-CLIL female students with 61.3% (Sylvén & Ohlander, 2014, p. 93). This descriptive difference between CLIL and non-CLIL students regarding form-meaning recognition knowledge has additionally shown to be significant ($p < 0.001$) with CLIL students outperforming non-CLIL students (Sylvén & Ohlander, 2014, p. 92). Moreover, male CLIL students achieved the highest score in all studies reviewed for this thesis, followed by Peters, Velghe and Van Rompaey (2019). Their study of 145 secondary school students in Belgium resulted in both the second highest and lowest percentage of correct answers given in the reviewed literature (Peters, Velghe & Van Rompaey, 2019, p. 755). While female 2nd year students reached the last place with a mean of 39.94% correct answers on the VocabLab Test (Peters, Velghe & Van Rompaey, 2019), male 4th year students achieved the second highest score with 73.61% correct answers on average (Peters, Velghe & Van Rompaey, 2019, p. 763). In comparison, all findings reported by other studies reviewed are

located somewhere between these two extreme scores, with participants either displaying scores of around 40% or between 50-60% on the form-meaning recognition tests used (Coxhead, Nation & Sim, 2015; Li & Kirby, 2015; Ordóñez et al., 2002, p. 721; Staehr, 2008). Both Staehr's (2008) 88 Danish secondary school students aged between 15-16 and Li and Kirby's (2015, p. 624) 246 EMI students in China achieved a mean of around 40% correct answers, with Staehr's (2008, p. 144) participants reaching 40.98% and Li and Kirby's (2015, p. 624) 42.85% correct answers on average, despite different form-meaning recognition tests used. Therefore, Staehr's (2008) and Li & Kirby's (2015) findings can be ranked rather at the lower end of the continuum, being higher than Peters, Velghe and Van Rompaey's (2019) results of female 2nd year students but lower than the results reported by other studies (Coxhead, Nation & Sim, 2015; Ordóñez et al., 2002). Even lower results regarding form-meaning recognition were found by Sylvén and Sundqvist (2016) in their validation study of the YLVAT, with 52 twelve year-old L1 Swedish participants reaching between 52.5 and 35% on average dependent on their gender. While the males outperformed the female participants significantly ($p=0.025$) regarding their receptive knowledge of the form-meaning link, no significant difference ($p=0.285$) could be found regarding written word recognition where boys reached a correct answer rate of 76.92% and girls 72.31%.

In their study of 88 4th and 5th grade bilingual secondary school students in the US Ordóñez et al. (2002, p. 720) detected a mean correct answer rate of 50.79% on the PPVT. Similarly, 227 secondary school students involved in Coxhead, Nation and Sim's (2015, p. 125) study in New Zealand displayed comparable results with the youngest participants, being 13 years old, achieving 54.02% correct answers on the VST and the oldest participant group aged 18 reaching 61.50% correct answers. Interestingly the highest score was not achieved by the oldest and therefore presumably most advanced group, but by the participants aged 17 with 66.81% correct answers (Coxhead, Nation & Sims, 2015, p. 125). Hence, a continuous growth of receptive form-meaning knowledge was only observed between the ages of 13 and 17, followed by a drop around the age of 18. Yet, since, the study was not longitudinal in the sense that the development of the same students was observed over several years, this trend might rather be attributed to the individual performance and proficiency of the age groups compared. This assumption is supported by several studies, such as Duff and Brydon (2020, p. 454), Makhoul's (2017, p. 1053) or Zichtl (2017, p. 65), who all report a significant increase of receptive vocabulary size over time. While Makhoul (2017, p. 1056) observed 600 participants in Israel for three years, Duff and Brydon (2020) and Zichtl (2017, p. 56) even cover a time span of five years, all finding a significant continuous improvement in receptive vocabulary size. Duff and Brydon (2020, p. 469) add that their 41 secondary school students improved on average by 460 word families per year. This corresponds to Webb and Chang's (2012, p. 116) findings when comparing 166 EFL learners in Taiwan, one group being an English focus class and the two other groups having no focus on the English language. In alliance to Duff and Brydon's (2020, p. 469) participants the English focus group

in Webb and Chang's (2012, p. 113) study increased by roughly 430 words a year, while participants in mainstream classes improved only slightly with a gain ranging between 114 - 176 words per year. Overall, Webb and Chang (2012, p. 120) emphasize that although an overall development could be observed, they detected considerable variation amongst the participants even within the same year of study. Moreover, they argue that not all learners' vocabulary learning paths are linear but that vocabulary learning rather comes in waves (Webb & Chang, 2012).

Concerning participants' performance regarding academic vocabulary knowledge at secondary school level, rather divergent findings were detected (Cameron, 2002; Cobb, 2000; Laufer, 1998; Townsend et al., 2016). The lowest result was reported by Laufer (1998, p.262), who states that the 26 10th and 22 11th grade secondary school students involved in her study in Israel managed to answer between 12.89% - 29.23% of the UWL items of the VLT (Nation, 1990; Nation, 1983) correctly. Conversely, the highest percentage of correct answers was also reported by a rather old study, namely Cobb (2000). In his study on form-meaning recognition knowledge of 750 Canadian secondary school students, he revealed that his participants could answer 74% of the K1 items and 68% of the UWL items of the VLT correctly (Cobb, 2000, p. 305). However, while Cameron's (2002) results on the Yes/No Test (Meara, 2015, p. 162) roughly approximate this result with 64% correct answers given, most other findings are clearly below this value. Townsend et al.'s (2016, p. 15) 256 participants reached a percentage of 61.39 correct answers on the Upper Level Spelling inventory (Bear et al. 2012) for instance, but they only managed to correctly identify 50.3% of the words of the self-developed Academic Spelling inventory. Additionally, Cameron (2002, p. 162) discovered that the 167 participants examined for her study could answer between 49-64% of the items of the AWL part of the Yes/No Test (Meara, 2015) but were only able to identify a correct meaning for 44-48% of the UWL items of the VLT. This again highlights the importance to differentiate measures of different word knowledge aspects more clearly and additionally shows how much influence academic vocabulary lists, which are consulted as a baseline for a test or study, have on the results.

While studies on the development of overall receptive vocabulary size are relatively frequent, as already seen above, only few studies specifically investigated the development of academic vocabulary. One exception is Luxton, Fry and Coxhead's (2017) study with over 2500 bilingual secondary school students in New Zealand. Using the academic section of the VLT, which covers words from the AWL, Luxton, Fry and Coxhead (2017, p. 18) discovered that over the timespan of six months academic vocabulary knowledge developed by 6.66-10.19 points out of a maximum score of 30. This equates to an improvement of 22.2-33.97%. Furthermore, a study by Coxhead and Boutorwick (2018) on 468 German secondary school students, combining native English speakers, highly proficient EFL learners and low proficiency EFL learners, provides more insight into the development of academic vocabulary. Using the VLT, Coxhead and Boutorwick (2018) assume that achievement of 26

out of 30 maximal points per frequency level covered in the test, can be equated to mastery of a level. While both native English speakers and proficient non-native speakers already mastered the K2 level when starting the study, low proficiency speakers needed one year to take this hurdle (Coxhead & Boutorwick, 2018, p. 597). Concerning academic vocabulary knowledge, native speakers took two years to master the AWL section of the VLT, while proficient EFL learners needed at least four years and low proficiency EFL speakers even took five years to master this section (Coxhead & Boutorwick, 2018, p. 598). However, all three groups of learners were able to develop their vocabulary significantly in all frequency levels at the beginning but reached a plateau at a certain level. This fossilization was first shown by native speakers, followed by proficient EFL learners and low proficiency EFL learners (Coxhead & Boutorwick, 2018, p. 599). A similar trend regarding overall language proficiency and academic vocabulary knowledge was detected by Townsend et al. (2016, p. 1), who reached the conclusion that orthographic knowledge of academic vocabulary has a significant impact on general language achievement. Additionally, Skjelde and Coxhead (2020, p. 12) report a medium relationship between 134 Norwegian secondary school students' academic form-meaning recognition knowledge and English course grades [$r_s=0.37, p<0.001$].

Zooming in on the influence of receptive vocabulary size on different language skills, Staehr (2008) found a strong relationship between receptive vocabulary size and three language skills (Staehr, 2008). With $r=0.83, p<0.01$ the correlation between receptive vocabulary size and reading comprehension was the strongest, followed by $r=0.73, p<0.01$ for writing and $r=0.69, p<0.01$ for listening (Staehr, 2008, p. 144). Moreover, a linear regression analysis showed that receptive vocabulary size was able to explain 72% of the variance in the reading test and at least 39% of the variance in the listening test (Staehr, 2008, p. 145). Realizing this possible relation several more research studies examined the relationship between receptive vocabulary size and reading performance (Li & Kirby, 2015, p. 627; Logan & Kieffer, 2021, p. 318), with similar results. Complementary to Staehr (2008) Li and Kirby (2015, p. 627) found that receptive vocabulary size significantly influences both performance in word reading and overall reading comprehension, which is also confirmed by Logan and Kieffer's (2021) study of 107 bilingual secondary school students. Building on these findings Duff and Brydon (2020) investigated how large the receptive vocabulary size gap between poor and successful readers was amongst 41 secondary school students. They found that poor readers knew around 1300 word families less than successful readers, pointing at a rather challenging but feasible target for learners to achieve (Duff & Brydon, 2020, p. 454).

Conversely to the studies above, which focused on the effect of receptive vocabulary size on different areas, other research studies examined potential factors that might influence receptive vocabulary knowledge. Firstly, several studies investigated the impact of different language learning approaches on receptive vocabulary size (Hahn, 2017; Luxton, Fry &

Coxhead, 2017; Mondria & Wiersma, 2018; Schwarz, 2020; Sundqvist, 2009; Webb & Chang, 2012; Zichtl, 2017). On a very general level, Luxton, Fry and Coxhead (2017, p. 19) detected that length of time spent in EMI schooling impacted the receptive vocabulary size of their participants. A similar result was reported by Webb and Chang (2012, p. 120), who found that students taking an English focus class achieved higher receptive vocabulary size scores than students without a focus on the English language. Furthermore, Mondria and Wiersma's (2018, p. 88-91) findings show that within a class language learning tasks consisting of both receptive and productive activities had the largest impact on vocabulary knowledge. Hence, EMI and language focus classes mostly consisting of a mixture of receptive and productive language learning tasks might provide the ideal input for receptive vocabulary size growth.

In addition to the impact of language learning tasks inside the classroom, other studies focused on the influence of Extramural English, being English encountered outside the walls of the classroom (e.g. Sundqvist, 2009). For a more detailed description and review of Extramural English see chapter 5. Overall, several studies confirmed the positive influence of involvement in Extramural English activities on receptive vocabulary size (Hahn, 2017; Schwarz, 2020; Sundqvist, 2009; Zichtl, 2017). While Hahn (2017, p. 82) found that amount of daily Extramural English involvement correlated significantly with performance on the Yes/No Test (Meara, 2015), $r=0.398$, $p<0.01$, regardless of the type of EE activity. Zichtl (2017, p. 82) adds to these findings, by reporting that watching English TV programs correlated strongest with scores on the VST with $r=0.282$, $p<0.05$, followed by reading in English with $r=0.279$, $p<0.05$, English speaking opportunities in spare time with $r=0.182$, $p<0.05$ and finally listening to English music with $r=0.162$, $p<0.05$. On a similar note, Peters, Velghe and Van Rompaey (2019, p. 766) developed a structural equation model with partly satisfactory fit indices examining the influence of various latent EE variables as well as biographical background variables on vocabulary knowledge. They found that EE involvement in playing online games had a strongly significant effect on vocabulary knowledge ($p<0.001$), which is again characterised by a strong gender difference (Peters, Velghe & Van Rompaey, 2019, p. 766). Regarding biographical background variables only the educational level of the students had a significant ($p<0.001$) effect on vocabulary knowledge. Additionally, educational level influenced involvement in written media ($p<0.05$) and other less wide spread EE actions ($p<0.05$). All other EE variables, such as English use with family and friends, involvement in written or audio-visual material with or without subtitles had no significant effect on receptive vocabulary knowledge (Peters, Velghe and Van Rompaey, 2019, p. 766). While Peters, Velghe and Van Rompaey's (2019), as well as Hahn's (2017) and Zichtl's (2017) research focus was on the overall receptive vocabulary size of participants, Sundqvist (2009) examined whether involvement in Extramural English has an impact on all frequency bands included in the VLT. Sundqvist (2009, p. 151) revealed that engagement in Extramural English activities had a positive impact on both K2 and K3 word

knowledge but not on the knowledge of vocabulary items found in the K5 frequency band. Consequently, Sundqvist (2009, p. 151) argues that Extramural English does not seem to have the same influence on word knowledge of all frequency bands and that more difficult and less frequent words appear to be less affected by extramural involvement.

Additional to effects of English contact in and outside the classroom biographical factors, such as age, length of instruction, bilingualism or gender were considered as potential predictor variables. Regarding the influence of age on receptive vocabulary size Coxhead, Nation & Sim (2015, p. 131) found that the mean score on the VST increased for their participants the older the pupils were. This implies that age does influence receptive vocabulary size positively, which can be connected to the influence of length of English instruction. Both Luxton, Fry & Coxhead (2017, p. 18) and Peters, Velghe and Van Rompaey (2019, p. 747) report that length of English instruction has a positive effect on receptive vocabulary size of participants, similar to the findings reported for age. Nonetheless, Zichtl (2017, p. 83) states that age of onset had no significant influence on the performance of her Austrian participants on the VST ($b = -326.749$, $\beta = -.116$, $t = -1.630$, $p > .05$), which although investigating a different concept to the studies before, appears to be connected, since early onset tends to go hand in hand with long English input. Hence, more studies investigating these separate but linked concepts should be conducted to examine the influence of educational and biographical variables on receptive vocabulary size more closely.

Regarding the impact of gender on receptive vocabulary size similar findings were detected by Peters, Velghe and Van Rompaey (2019, p. 764) and Sundqvist (2009, p. 150), both reporting a significant gender difference with boys outperforming girls. Finally, Luxton, Fry and Coxhead (2017, p. 18) touch on the extensive field of bilingualism and multilingualism by examining whether the home language spoken and the main educational language spoken have a significant influence on receptive vocabulary size. Investigating secondary school students in New Zealand, partly speaking Maori at home or as educational language, they detected that both factors have an impact on participants' receptive vocabulary size. Hence, to summarize biographical background variables appear to be connected to receptive vocabulary size and should therefore be further investigated in future studies.

4.1.3 Tertiary education

Comparing receptive vocabulary size estimates of secondary school students to those of participants involved in tertiary education reveals several similarities. Similar to the results found for secondary school, receptive vocabulary size estimates at tertiary level span a rather large range from 7274.75 word families (Yang et al., 2019, p. 851) to 1226 words (Nurweni & Read, 1999, p. 168). Moreover, receptive vocabulary size does not seem to change considerably from secondary to tertiary level education, since estimates are rather comparable. In between the two extreme scores many receptive vocabulary size estimates result in values between 3000 to 4000 words known (Miralpeix, 2007; Roghani & Milton,

2017; Schmitt & Meara, 1997; Yamamoto, 2014). A rather recent study by Roghani and Milton (2017, p. 152) reveals that students of a foreign language teaching school in Iran knew around 3000 words receptively. This rather vague estimate is supported by the findings of Miralpeix (2007) who discovered that 64 intermediate EFL students in Spain reached an average receptive vocabulary size of 3950 words based on the VLT (Miralpeix, 2020, p. 197). Comparable results are additionally reported by Schmitt and Meara (1997, p. 23) in a longitudinal study, who report that 86 English major students in Japan reached a receptive vocabulary size of 3900 words at the beginning of his study and 4230 words after nearly a year of study, depicting the increase of receptive vocabulary size throughout tertiary level education. This is additionally highlighted by Miralpeix and Muñoz's (2018), Ozturk's (2016) and Miralpeix's (2007) findings involving more advanced EFL learners at tertiary level all achieving a receptive vocabulary size between 5000 to 6000 words. While Miralpeix and Muñoz's (2018, p. 12) 42 Spanish English major students reached a receptive vocabulary size of 5100 words, Ozturk's (2016, p. 9) 174 undergraduate students in Turkey achieved 5705.17 words and Miralpeix's (2007) 93 advanced EFL students 5954 words.

Additional to studies indicating a total receptive vocabulary size for their participants, several researchers present their results in percentages of correct answers on the test method used (e.g. Ma & Lin, 2015; Staehr, 2009), which show a large diversity. While 124 college students in Taiwan achieved the highest score with 88% correct answers on the VLT (Ma & Lin, 2015, p. 7), 174 undergraduate students in Turkey received the lowest score of the studies reviewed with 47.05% correct answers on a shortened version of the VST (Ozturk, 2016, p. 9). Similar to Ma and Lin's (2015) result Staehr (2009, p. 590) detected a percentage of 82.44% correct answers on the VLT by 115 Danish tertiary level students. These findings are set apart from other findings by a small gap, with the next highest correct answer rates found 74% correct answers given by 55 1st year and 45 4th year students in Turkey (Ozturk, 2015, p. 99) and 72% of the VLT items answered correctly by 150 1st tertiary level students in China (Zhang & Lu, 2015, p. 745). Moreover, several studies report a correct answer rate of 60-70% (Fan, 2000; Lee, 2020; Qian, 2002; Wang, 2014). Lee (2020, p. 7) states that 492 students in Korea were able to answer 69.9% of the VLT items correctly on average. Congruent to this result Qian (2002, p. 528) identifies a correct answer rate of 66.6% on the VLT amongst his 217 intermediate and advanced Canadian tertiary level students. Comparable results were found by Wang (2014, p. 2461) with 64.44% correct answers on the VLT for 168 non-English major students in China, and by Fan (2000, p. 111) with 63.51% of the items answered correctly by 138 Chinese students. The lowest results were found for two recent studies conducted in Iran (Dabbagh & Enayat 2019; Vafaei & Suzuki, 2020) and in Turkey (Ozturk, 2016). Dabbagh and Enayat (2019, p. 580) reports that 67 English major students were only able to answer 56.25% of the VLT items correctly. An even lower percentage has been detected by Vafaei and Suzuki (2020, p. 398) with 49.50% correct answers on an aural version of the VST. While it could be argued that Vafaei and

Suzuki's (2020) results could be partly explained by added difficulty through the aural presentation of the test items, Ozturk (2016, p. 9) used a written short version of the VST and still achieved the lowest score of the studies reviewed.

Regarding receptive form-meaning knowledge of academic words rather few studies have been conducted (Masrai & Milton, 2018; Nurweni & Read, 1999; Ozturk, 2015; Townsend et al., 2012; Tschirner, 2004), which have resulted in divergent findings. While Ozturk (2015, p. 99) reports that his participants were able to answer 85% of the AWL part of the VLT correctly at the first data collection and knew even more words (88%) at the second data collection, Nurweni and Read (1999, p. 168) detected a rather low value of 29.58% of the academic words known. Moreover, Tschirner (2004, p. 31) states that out of his 142 1st semester students in Germany 76% answered over 80% of the academic items correctly and 47% even reached a score of over 90% correct answers. Townsend et al. (2012, p. 509) differentiates these results a bit more by showing that students with increased English contact performed better on academic word knowledge (22.70 points) than students with less English contact (19.68 points) and that students with high socio-economic status (23.77 points) had an advantage in comparison to students with a rather low socio-economic status (19.39 points).

While studies estimating receptive vocabulary size are more frequent at tertiary level than at secondary school level, literature appears to be rather short on studies focusing on the development of vocabulary knowledge at tertiary level. One of the exceptions is Yamamoto's (2014, p. 234) study investigating the receptive vocabulary size of 22 Japanese 1st year students using the VLT. Using a repeated measures t-test Yamamoto (2014, p. 236) found a significant increase of receptive vocabulary size from the 1st to the 2nd data collection, $t(21) = 2.75$, $p < 0.0167$, $d = 0.60$. However, the validity of the performed statistics needs to be questioned, since the data sample appears to be very small for inferential statistics and no information about assumptions, such as normal distribution is given. Nonetheless, Ozturk (2016, p. 9) also detected a significant development of receptive vocabulary size amongst 174 tertiary level students, $F = 11.128$, $p = 0.000$. Schmitt and Meara (1997, p. 17) specified that their participants showed an increase of 330 words in approximately one year, which was a significant increase on all frequency levels observed. A similar positive result was found by Gyllstad (2004), who investigated the receptive form-meaning recognition knowledge of 90 Swedish university students at three different proficiency levels. While no significant difference in the scores of groups A and B could be found, a significant difference ($p < 0.05$) was detected between groups A and C as well as groups B and C. This is also visible in the percentages of correct answers, which ranged from 57.67% over 57.92% to 67.3%.

Different findings were reported by Ozturk (2015, p. 97), who also used a repeated measures t-test to investigate the development of the receptive vocabulary size of 17

students in Turkey, which again appears to be a rather small sample size for the statistics performed. Nevertheless, Ozturk (2015, p. 98) reports that he could not find any statistically significant growth of receptive vocabulary size between the 1st year and the 4th year of studies. However, one needs to take into account that this study is pseudo-longitudinal, comparing different individuals to each other. Hence, differences or no differences in performance might be attributed to variation regarding the proficiency of the participants compared rather than evidence for (no) vocabulary development.

Similar to studies at the secondary level a relatively large number of studies at the tertiary level appears to be interested in the influence of receptive vocabulary size on other language skills. Studies such as Miralpeix and Muñoz (2018), Wang (2014) and Roche and Harrington (2013) investigated the influence of receptive vocabulary size on overall language proficiency, finding significant relationships between the two variables. While Wang (2014, p. 2462) states the overall language proficiency was moderately [$F=106.086$, $p<0.001$, $r^2=0.39$] related to receptive vocabulary size measured by the VLT (Schmitt, Schmitt & Clapham, 2001) for his 168 Chinese tertiary level students, Miralpeix and Muñoz (2018, p. 13) found a strong correlation [$r=0.618$, $p<0.001$] between receptive vocabulary size measured by the X-Lex (Meara, 2005) and overall proficiency. Moreover, Miralpeix and Muñoz's (2018, p. 14) research reveals that 36.6% of the total variance of overall proficiency can be attributed to receptive vocabulary size.

Narrowing the focus from overall proficiency to academic performance, Roche and Harrington (2013) found a significant connection between receptive vocabulary size and overall academic achievement for 70 tertiary students tested with the help of a Yes/No test (Roche & Harrington, 2013, p. 9). This is confirmed by Masrai and Milton (2018, p. 50), who found a strong positive correlation [$r=0.73$, $p<0.001$] between test performance and overall academic performance. Townsend et al. (2012, p. 511) expands on these findings by claiming that between 2-7% of the variance within academic performance can be explained by receptive vocabulary size. Szabo, Stickler and Adinolfi (2021, p. 1538) even expand on this by stating that form-meaning recognition knowledge measured with the VLT explains over 30% of the variance in academic achievement.

Additional to studies with a focus on overall language proficiency several studies have examined the influence of receptive vocabulary size on individual language skills, such as reading, writing, listening or speaking (e.g. Dabbagh & Enayat, 2019; Miralpeix & Muñoz, 2018; Yang et al., 2019). Concerning receptive skills, reading has been researched the most, with all studies finding a significant connection between the two abilities (Ehsanzahdeh, 2012; Ma & Lin, 2015; Miralpeix & Muñoz, 2018; Qian, 2002). For instance, Ehsanzahdeh (2012, p. 34) has found a strong correlation between inferencing success and performance on the VLT for 33 undergraduate students in Iran [$r=0.76$, $p<0.001$], similar to Miralpeix & Muñoz's (2018, p. 13) finding a correlation strength of $r=0.519$, $p<0.01$, which is smaller but

still a medium correlation. Furthermore, Miralpeix & Muñoz' (2018) results on the influence of receptive vocabulary size on reading proficiency are comparable to Ma & Lin's (2015, p. 9) findings with 25.1% of the total variance in reading performance explained by receptive vocabulary size opposed to 20%. However, Qian (2002, p. 529) states that for academic reading receptive vocabulary size additionally to receptive vocabulary depth could explain even over 50% of the variance. Hence, it remains to be seen whether different reading genres are influenced differently by receptive vocabulary size knowledge.

Another receptive language skill, namely listening, was also researched by Miralpeix and Muñoz (2018, p. 13), who found that listening was at least moderately connected to receptive vocabulary size with $r=0.420$, $p<0.01$. Similar findings were reported by Cheng & Matthews (2018, p. 16) and Staehr (2009, p. 591), who even reports a correlation coefficient of $r=0.70$, $p<0.01$ for his 115 Danish participants. Yet, the strength of this connection partly contradicts the findings of Miralpeix and Muñoz (2018, p. 13) in whose study listening was the least connected skill to receptive vocabulary size. Nonetheless, this difference might again be explained by the methodology used, namely the X-Lex on the side of Miralpeix and Muñoz (2018) and the VLT on the side of Staehr (2009), which measure two different word knowledge aspects.

Regarding a potential connection between receptive vocabulary size and productive language skills, research on writing is clearly favoured, finding several significant correlations of varying size (Dabbagh & Enayat, 2019; Roche & Harrington, 2013; Yang et al., 2019). Furthermore, Dabbagh and Enayat (2019, p. 580) report that receptive vocabulary size was able to explain at least 19% of the variance within the writing performance of the 67 Iranian students taking the VLT. In comparison only one study, namely Miralpeix & Muñoz (2018, p. 13), is concerned with the influence of receptive vocabulary size on speaking, finding a significant moderate correlation.

While there appears to be a clear connection between receptive vocabulary size and several language skills, it still remains questionable how receptive vocabulary size itself can be influenced at tertiary level. Several research studies (e.g. Fan, 2000; Yamamoto, 2014; Zhang & Lu, 2015) investigated the impact of learning strategies on receptive vocabulary size and have continuously found positive results. For example, Yamamoto (2014, p. 238) states that systematic learning strategies has a positive influence on the receptive vocabulary size of 22 Japanese students. Moreover, Zhang & Lu (2015, p. 748) found that vocabulary learning strategies based on word form learning had a large explanatory power, providing explanation for 31.4% of the total variance of receptive vocabulary size of 150 Chinese 1st year students. These were followed by visualization strategies with an explanatory power of 13.7%, training of association knowledge, such as synonyms or antonyms with 9.3%, repetition with 7.4% and lastly list learning with 5.8%. Moreover, Peters's (2014, p. 75) study of 37 Belgian tertiary students already appears to support Zhang and Lu's (2015)

findings regarding the impact of repetition. Another interesting aspect covered by the literature is the potential of productive and receptive learning for receptive vocabulary size. While Tschirner (2004, p. 33) states that out of the 142 German students participating in his study those students being more involved in productive learning were also more successful in productive vocabulary knowledge and students preferring to learn receptively displayed the larger receptive vocabulary size, Webb (2009) reaches a somewhat different conclusion. Using an experimental design, Webb (2009, p. 360) finds that 62 Japanese tertiary level students benefited more from productive learning than from receptive learning, since productive learning led to more productive and receptive vocabulary knowledge.

Summarizing these findings with regard to this study reveals that studies largely appear to agree that form-meaning recognition knowledge develops over time at all educational levels. In comparison, vocabulary size estimates are quite divergent at all educational stages. On average participants at secondary school level answer between 40%-60% of test items correctly, while at tertiary level the correct answer rate increases to 60%-70%. Consequently, a higher number of correct answers can be expected for this study observing tertiary level students majoring in English.

Regarding knowledge of academic vocabulary findings are similarly divergent across educational levels as for general vocabulary knowledge. This might be partially due to the choice of target items investigated or test instruments used. Nonetheless, studies across all educational levels agree on the existence of a connection between form-meaning recognition knowledge and overall language proficiency. Furthermore, form-meaning recognition knowledge is consistently positively influenced by increased English contact through EMI or EE, as well as some biographical variables.

4.2 What we already know about form-meaning recall knowledge

As suggested by Figure 4.5 below productive vocabulary size appears to have aroused the interest of researchers later than receptive vocabulary size. Although this literature review does not claim to be fully comprehensive but compares and contrasts 28 of the most central research studies in the field, no research studies with this focus, published between 2001 and 2008 as well as 2010 and 2013, were included, while more recent studies from 2014 onwards were found.

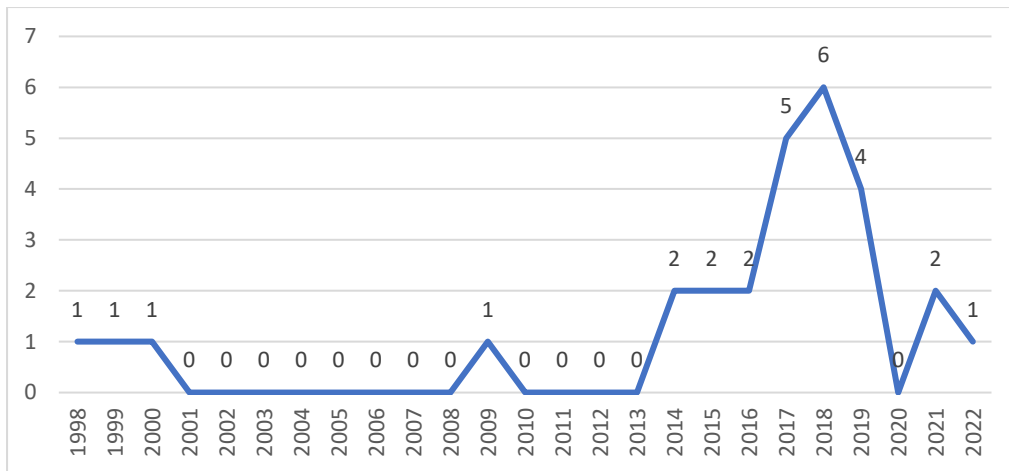


Figure 4.5: Distribution of publication year of form-meaning recall studies

As Figure Figure 4.6 shows, one similarity between research on productive and receptive vocabulary size can be detected which is related to the geographical distribution of the studies². Both types of vocabulary size appear to be popular research foci in Asian countries, such as China (e.g. Cheng & Matthews, 2018; Fan, 2000; Li & Hafner, 2021; Sun et al., 2018; Yang et al., 2019; Zhong, 2018), Japan (e.g. Fitzpatrick & Clenton, 2017; Ishihara, Okada & Matsui, 1999; Uchihara & Saito, 2019; Yamamoto, 2014) or Korea (e.g. Lee, 2019a; Lee, 2019b). Furthermore, many other studies appear to be distributed in European countries, such as Austria (Ghamarian, 2022; Zichtl, 2017), Spain (Miralpeix & Muñoz, 2018), Netherlands (Mondria & Wiersma, 2018), Sweden (Olsson & Sylvén, 2015; Sylvén & Sundqvist, 2016), Germany (Schuth, Köhne & Weinert, 2017) and Turkey (Ozturk, 2015).

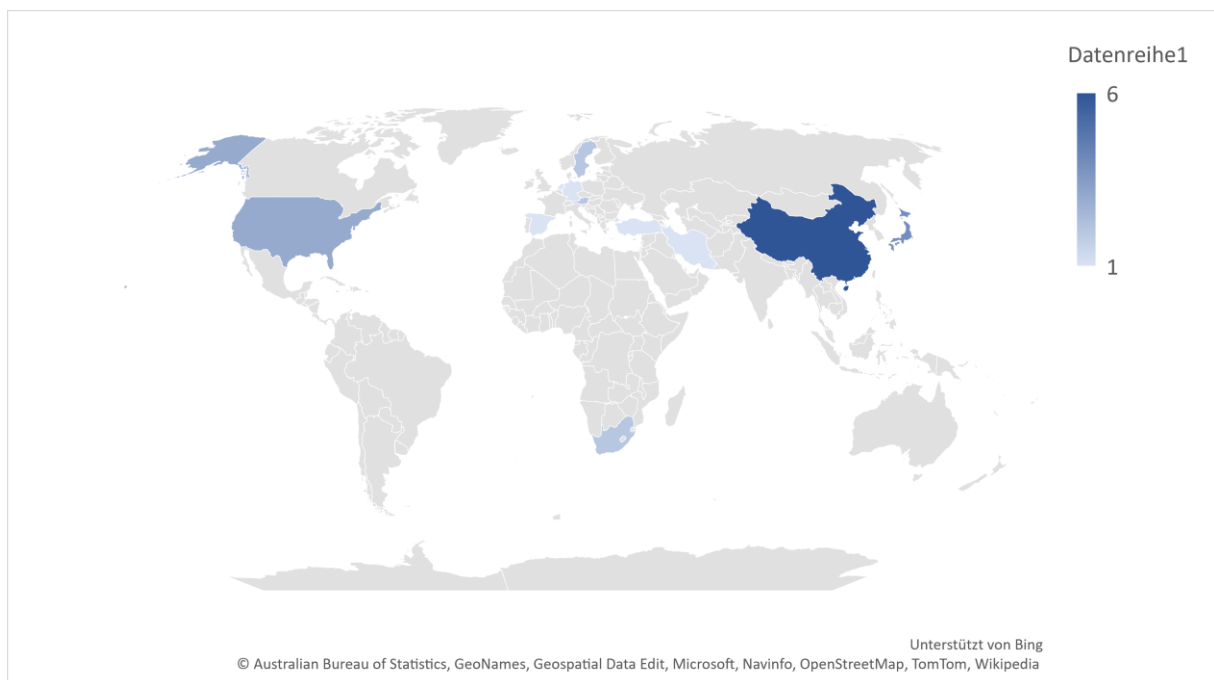


Figure 4.6: Geographical spread of form-meaning recall studies

² A more detailed map can be found in appendix 4.2.

Yet, productive vocabulary size studies located in the US (e.g. Mc Keown et al., 2018; Pham & Khonert, 2014; Vagh, Pan & Mancilla-Martinez, 2009) appear to be much rarer than receptive vocabulary size studies. Instead, the research gap located for receptive vocabulary size in African countries is less striking for productive vocabulary size, with at least two studies being conducted on this continent (e.g. Asaad & Shabdin, 2021; Nizonkiza, 2016).

Regarding the distribution of methodological tools used to investigate productive vocabulary size a comparable picture to receptive vocabulary size can be found in Figure 4.7.

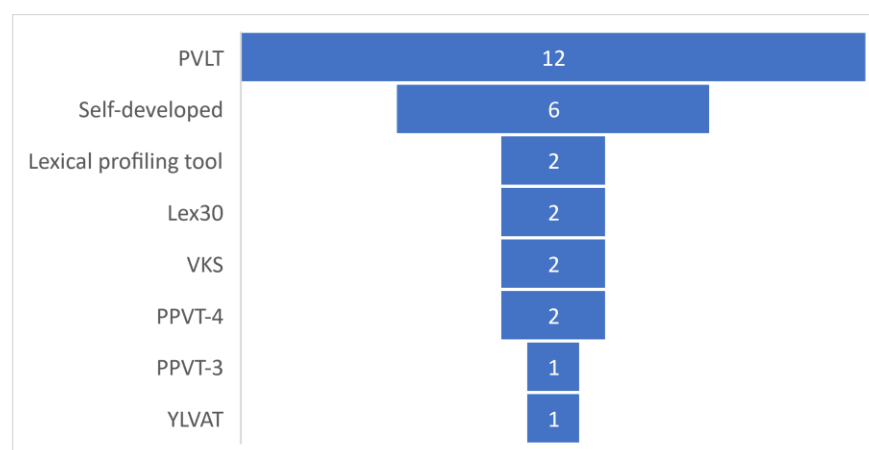


Figure 4.7: Distribution of testing method in form-meaning recall studies

Similar to an observable preference for the VLT (e.g. Nation, 1990; Schmitt, Schmitt & Clapham, 2001; Webb, Sasao & Ballance, 2017) as testing method for receptive vocabulary size, the PVLt (e.g. Cobb, 2000; Laufer & Nation, 1999) appears to be favored for investigations of productive vocabulary size (e.g. Asaad & Shabdin, 2021; Cheng & Matthews, 2018; Fan, 2000; Laufer, 1998; Lee, 2019a; Lee, 2018; Makhoul, 2017; Ozturk, 2015; Yamamoto, 2014; Yang et al., 2019). This is directly followed by the use of self-designed tests (e.g. Li & Hafner, 2021; Ishihara, Okada & Matsui, 1999; Miralpeix & Muñoz, 2018; Mondria & Wiersma, 2018; Nizonkiza, 2016), which can be partly explained by the comparable small number of available tests for productive vocabulary size. However, few studies (e.g. Ghamarian, 2022; Olsson & Sylvén, 2015) decided for the application of lexical profiling tools, such as AntWordProfiler (Anthony, 2022), which are used to investigate the vocabulary of written texts or for other existent tests as the Lex30 (Fitzpatrick & Clenton, 2017; Uchihara & Saito, 2019), the VKS (Tahmasbi & Farvardin, 2017; Zhong, 2018) or one of the PPVT versions (Pham & Khonert, 2014; Sun et al., 2018; Vagh, Pan & Mancilla-Martinez, 2009).

Concerning the coverage of different educational levels, a similar picture than for receptive vocabulary size can be found with the least studies conducted at primary school level, followed by secondary school level and most studies conducted at tertiary level (see Figure 4.8).

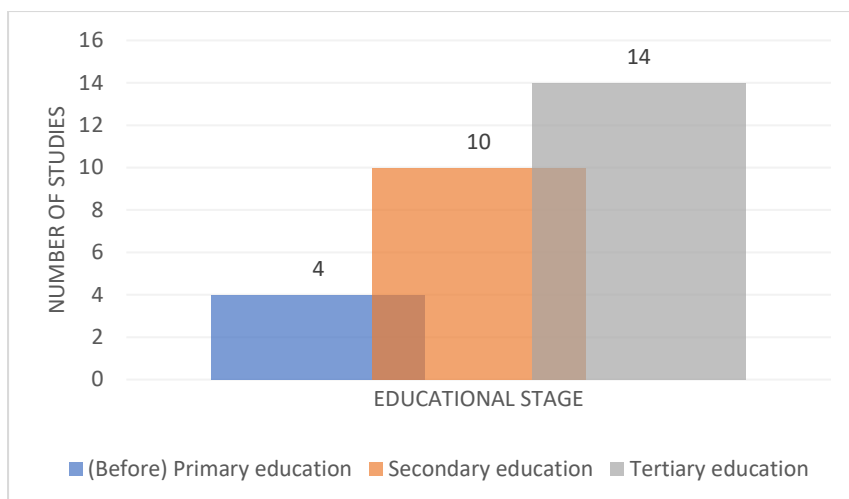


Figure 4.8: Educational level of participants involved in form-meaning recall studies

In the following research findings regarding the productive vocabulary size of primary and secondary school pupils will be reviewed, followed by a summary of the most important findings at tertiary level.

4.2.1 Primary and secondary education

Starting with the studies focusing on pre-school children Sun et al. (2018) detected similar productive vocabulary size growth as Vagh, Pan and Mancilla-Martinez (2009, p. 1550) for monolingual children before school entry in the US. While Sun et al. (2018, p. 16) report that 43 pre-school children in China significantly ($p < 0.001$) developed their English productive vocabulary size by 9.6 points on the PPVT-4, Vagh, Pan and Mancilla-Martinez (2009) found an increase of 10.5 points on the PPVT-three between the age of 30 and 36 months. In contrast, Pham and Khonert (2014, p. 775) found larger gains for 33 primary school children with L2 English also located in the US. The children started with a correct answer rate of 28% on the Expressive One Word Picture Vocabulary Test (Michalec & Henninger, 2011) and moved over 35% to 50% and finally 64% correct answers on the test, which equates a growth of at least 36% (Pham & Khonert, 2014, p. 775). Interestingly, Vagh, Pan and Mancilla-Martinez (2009, p. 1550) also approach another interesting topic by reporting that in contrast to the development of monolingual children included in their study, who reached a mean score of 16.9 points at the first data collection and 27.4 points at the second data collection, bilingual children achieved much higher scores of 71.5 and 72.9 points on the PPVT-3. Hence, bilingual children appear to have a considerably larger productive vocabulary size at this age according to their study, while monolingual children show a greater development over time (Vagh, Pan & Mancilla-Martinez, 2009). Yet, these results need to be treated with caution since further investigations regarding the influence of multilingualism on productive vocabulary size need to be conducted.

While results regarding productive vocabulary size appear to be at least roughly comparable at primary school level, findings at secondary school level remain difficult to interpret due to

different counting units applied by researchers. In a pseudo-longitudinal study including 26 10th and 22 11th graders Laufer (1998, p. 262) reports that the two groups differed concerning their productive vocabulary size by 11.68 points, which was a significant difference. Although Fitzpatrick and Clenton (2017, p. 852-854) also report productive vocabulary size in mean points achieved on the Lex 30 they only state that 80 secondary school students participating in a first study reached a mean score of 38.5 points and a second group of 100 pupils achieved only 32.5 points. Sylvén and Sundqvist (2016, p. 9) report in their validation study of the YLVAT that their 52 twelve year old L1 Swedish participants were able to answer between 30.83 and 41.6% of the twelve items taken from the PVLIT correctly on average dependent on their gender. While many of these studies individually give interesting insights, all but Sylvén and Sundqvist (2016) remain incomparable, since they give no maximum score for the tests used, which hinders interpretation and comparison of the scores in percentages across studies.

Regarding productive vocabulary size development, the studies of Makhoul (2017) and Tahmasbi and Farvardin (2017) report contradictory results. Makhoul (2017, p. 1059) detected no significant growth of productive vocabulary size over time, while Tahmasbi and Farvardin (2017, p. 5) found a significant effect for time ($F_{2, 248} = 726.611, p = .000$, partial $\eta^2 = 0.854$). However, while Makhoul (2017) focused on general receptive and productive vocabulary growth, Tahmasbi and Farvardin (2017) further examined the influence of different task types on receptive and productive vocabulary development. Investigating the productive vocabulary size of 144 secondary school students with the help of the VKS Tahmasbi and Farvardin (2017, p. 5) found that the task type used for learning had a significant effect on productive vocabulary size [$F_{2, 124} = 9.985, p = 0.000, \eta^2 = 0.287$]. More concretely participants who practised vocabulary knowledge through a paragraph writing task, which required them to use vocabulary items in context, outperformed pupils who only integrated words into single sentences and students who needed to order given sentence parts correctly additionally to students involved in translation tasks (Tahmasbi & Farvardin, 2017, p. 6-7). Using a similar experimental approach Mondria & Wiersma (2018, p. 88) examined which influence receptive, productive or mixed learning had on the productive vocabulary size development of 198 secondary school pupils in the Netherlands. At the immediate post test administered directly after the tasks were completed pupils who had learned productively achieved the highest score with 96.9% correct answers on the self-designed translation test. They were followed by the mixed modality group with 88.12% correct answers and lastly by pupils learning merely receptively with 52.5% correct answers on the test (Mondria & Wiersma, 2018, p. 91). This result could be directly transmitted to the delayed post test conducted some time after the learning tasks, with pupils in the productive learning group still recalling 25.63% of the target words, mixed modality pupils knowing 20.63% and the receptive learning group only recalling 11.9% of the words (Mondria & Wiersma, 2018, p. 91). Hence, it seems that task type has a significant influence

on productive vocabulary size development and productive learning tasks appear to be more beneficial for productive vocabulary size in comparison to receptive or mixed tasks.

Investigating the influence of out of class English exposure on productive vocabulary use of academic words in 525 student texts, Olsson & Sylvén (2015, p. 93) found that male CLIL students, who were involved the most in extramural English activities also used the highest proportion of academic words in their essays. Yet, they displayed no advanced progress regarding the use of academic vocabulary over time in comparison to other participants (Olsson & Sylvén, 2015, p. 94). Hence, Olsson & Sylvén (2015, p. 94) conclude that extramural English exposure does not seem to have a considerable impact on the development of academic vocabulary use over the whole time of the investigation. Instead, they found that total frequency of extramural English exposure correlated significantly [$r=0.25$, $p=0.007$] with the proportion of academic vocabulary used in the first assignment but did not correlate with later written assignments (Olsson & Sylvén, 2015, p. 93). This validates the findings of Schwarz (2020, p. 247) at a more limited scope, who could not find a significant connection between overall productive vocabulary size and extramural English exposure. Nevertheless, Zichtl (2017, p. 84) managed to find a connection between productive vocabulary size and extramural English activities. Therefore, it remains to be seen which criteria influence the impact of extramural English exposure on both overall productive vocabulary size and form-meaning recall knowledge of academic words.

Regarding the influence of explicit instruction on academic form-meaning recall knowledge McKeown et al. (2018, p. 591) argue that explicit instruction seems to have a positive impact on performance. By investigating 105 middle school students, divided into 62 treatment group students receiving explicit instruction and 43 control group students receiving no instruction, they found that the treatment group significantly outperformed the control group regarding academic form-meaning recall knowledge [$t(102) = 7.04$, $p<0.001$] (McKeown et al., 2018, p. 591). However, the experiment showed that the influence of explicit instruction ceased slightly over time. While there was still a significant difference between the treatment group and the control group after one year [$F(1, 36) = 8.31$, $p= .007$], both groups showed no significant development of academic form-meaning recall knowledge anymore (McKeown et al., 2018, p. 599).

Concerning the influence of academic form-meaning recall knowledge on overall language proficiency, Schuth, Köhne and Weinert (2017, p. 160) report interesting results for 173 4th graders in a German elementary school. Using regression modelling, the researchers revealed that academic form-meaning recall knowledge significantly predicts reading [$\beta=0.43$, $p<0.001$] and writing performance [$\beta=0.39$, $p<0.001$] as well as grades on social studies [$\beta=0.51$, $p<0.001$] and maths [$\beta=0.37$, $p<0.001$] (Schuth, Köhne & Weinert, 2017, p. 162).

4.2.2 Tertiary education

Similar to estimates of receptive vocabulary size at tertiary level, results concerning productive vocabulary size are reported either as total scores (Cheng & Matthews, 2018; Lee, 2019b), percentage (Fan, 2000; Miralpeix & Muñoz, 2018; Nizonkiza, 2016; Uchihara & Saito, 2019) or number of word families known (Yamamoto, 2014; Yang et al., 2019). Yamamoto (2014) and Yang et al. (2019) indicate a total productive vocabulary size. While Yamamoto (2014, p. 235) states that 22 1st year students at a university in Japan have an average productive vocabulary size of 1306 word families, Yang et al. (2019, p. 851) recorded a larger average productive vocabulary size of 3070.34 word families for 59 Chinese students, both using the PVLТ as measurement instrument. This difference can however be attributed to the fact that Yang et al.'s (2019) participants were graduate students, hence displaying a larger productive vocabulary size than Yamamoto's (2014) 1st year students.

Taking a look at the correct answer rates of participants rather similar findings can be reported, all ranging between 40-70% correct answers given (Fan, 2000; Miralpeix & Muñoz, 2018; Nizonkiza, 2016; Uchihara & Saito, 2019). The most recent study by Uchihara & Saito (2019, p. 5) conducted in Japan revealed that 39 1st year students were able answer 42.57% of the Lex 30 test correctly. Similar, Nizonkiza (2016, p. 178) discovered that 50.73% of all items could be answered correctly by 204 university students in South Africa. The highest result reported can be found in Fan's (2000, p. 114) study of 138 tertiary level students in China, who answered 69.4% of a shortened version of the PVLТ correctly. In a pseudo-longitudinal study conducted by Ozturk (2015, p. 96-97) in Turkey the productive vocabulary size of 55 1st year students is compared to the productive vocabulary size of 45 4th year students. Overall, the results appear to suggest a significant increase of productive vocabulary size, with 1st year students answering only 40% of the PVLТ correctly and 4th year students answering at least 44% of the items correctly.

Similar findings are reported for academic word knowledge only, which increased from 40% to 43% correct coverage (Ozturk, 2015, p. 99). Studies by Yamamoto (2014) and Ghamarian (2022) support these findings, both using a longitudinal design. Investigating 22 Japanese university students Yamamoto (2014, p. 236) has found a strong significant development of productive vocabulary size within the time of observation [$t(21)=3.42$, $0<0.0034$, $d=0.75$]. Similarly Ghamarian (2022, p. 287) reports that the proportion of academic [$F(2, 70) = 27.239$, $p < .001$] and technical/ low-frequency vocabulary [$F(2.00, 70.00) = 67.368$, $p < .001$] in student texts increased significantly over the time span of three years, while high-frequency vocabulary decreased significantly [$F(1.796, 64.639) = 107.154$, $p < .001$; $F(1.725, 62.084) = 9.133$, $p = .001$]. Yet interestingly the percentage of technical/low frequency vocabulary used within the texts already increased significantly after half a year from 16.19% of the overall tokens being technical words to 19.65% (Ghamarian, 2022, p. 286). In

contrast, academic vocabulary only increased significantly after the whole observation time. Hence, it will be interesting to see whether the results concerning academic vocabulary knowledge can be validated in this study using different methodological tools.

Comparable to studies at secondary school level, several studies examined potential beneficial factors for productive vocabulary size development (Fan, 2000; Lee, 2019b; Lee, 2018; Li & Hafner, 2021; Yamamoto, 2014). Both Yamamoto (2014, p. 238-239) and Fan (2000, p. 115) report that application of vocabulary learning strategies has a positive effect on productive vocabulary size. In more detail, Fan (2000, p. 115) examined the use of seven different vocabulary learning strategies, finding a significant connection of productive vocabulary size to all of them. Furthermore, Lee (2019b), Lee (2018) and Li and Hafner (2021) investigated the influence of extramural English exposure on productive vocabulary size at tertiary level. Opposed to the mixed findings detected at secondary school level, all studies at tertiary level suggest that involvement in extramural English activities has a positive impact on productive vocabulary size. In more detail, Lee (2019b, p. 120) states that the more varied extramural English exposure of the 94 Korean participants was, the better their productive vocabulary size was [$r=0.51$, $p<0.01$]. While both studies by Lee focus on different extramural English activities Li and Hafner (2021, p. 9) concentrated on the effect of mobile app learning on participants' productive vocabulary size. In contrast to a group learning with classical word cards, pupils using mobile apps for learning clearly outperformed the other group regarding productive vocabulary knowledge [$t(83)=2.59$, $p=0.011$, $d=0.56$]. Therefore, it seems that engagement in out of school English activities appears to have a positive effect on productive vocabulary size at tertiary level. However, a lack of studies focusing on the influence of extramural English on specific vocabulary types, such as academic vocabulary, can be detected. Consequently, this study will work towards filling this gap in present-day research.

To sum up less studies on productive form-meaning knowledge have been conducted in comparison to the receptive counterpart. Consequently, information about the development of productive form-meaning knowledge is scarcer, with not all studies observing a significant development of form-meaning recall knowledge in the times observed. Similarly form-meaning recall knowledge reveals a broader range of correct answers with 40%-70% correctly answered items at tertiary level. Moreover, divergent to findings on form-meaning recognition knowledge, form-meaning recall knowledge appears to be rather positively influenced by productive learning tasks in class, such as writing texts, only then being followed by a mix of tasks. In contrast, studies on form-meaning recognition knowledge rather reported mixed tasks to have the largest effect. Concerning out-of-class activities studies reveal mixed findings regarding the effect of EE on form-meaning recall knowledge.

4.3 What we already know about multiple meaning recognition knowledge

Regarding polysemous word knowledge 19 research studies were reviewed, out of which most projects were conducted rather recently. As visible in Figure 4.9, the oldest study by Schmitt was conducted in 1998 and the most recent study was published in 2022 (Söderström, Thorén & Karlsson, 2022).

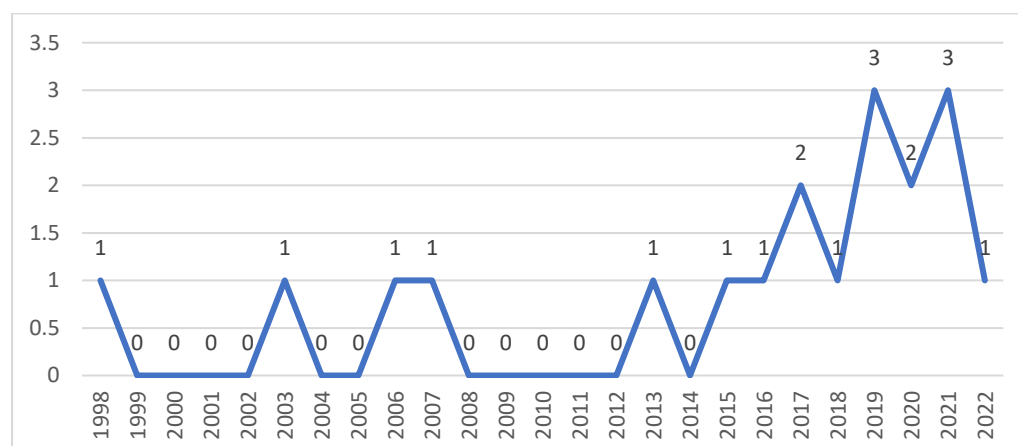


Figure 4.9: Distribution of publication year of multiple meaning recognition studies reviewed

In other words, it seems as if polysemous word knowledge has been established as a currently relevant topic in the research field. This is additionally supported by the number of different research instruments used in the different research projects considered (see Figure 4.10). This exemplifies that at the moment no standardized tests, which are normally favored by researchers, have been developed.

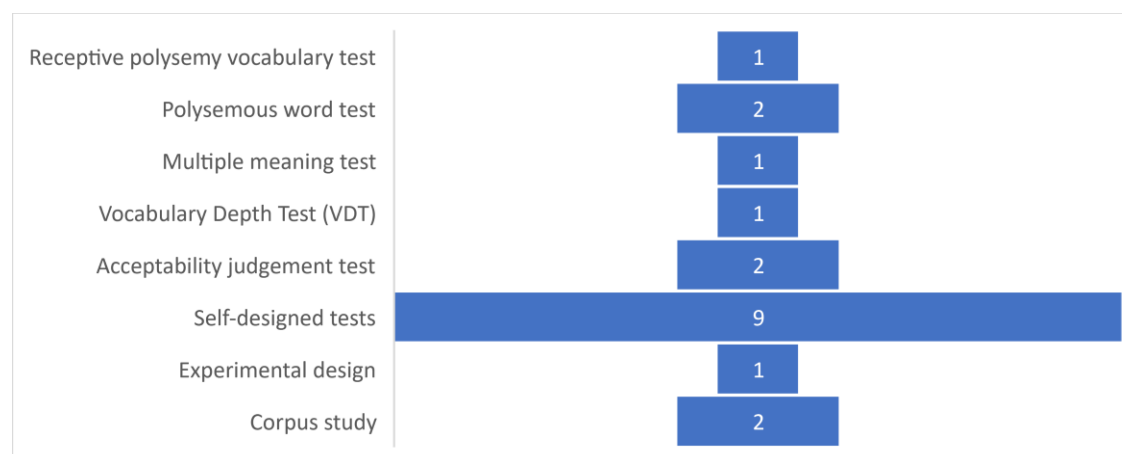


Figure 4.10: Distribution of testing method in multiple meaning recognition studies

While self-designed tests, developed by the researchers for the specific aim of their own studies, are the most popular methodological option chosen, seven different research instruments for polysemous word knowledge were applied. Out of the different types of vocabulary tests used, the Polysemous word test (Logan & Kieffer, 2017) and acceptability judgement tests (Maby, 2016; Zhang & Wen, 2019) were most popular, followed by the Receptive polysemy vocabulary test (Booton et al., 2021), the Multiple meaning test (Qian,

2002) and the Vocabulary depth task (Richard, 2011). Moreover, corpus based and experimental approaches were taken to examine polysemous word knowledge.

Regarding the regional spread of studies on polysemous word knowledge, many studies were conducted in L1 English countries, such as the USA, Canada or Great Britain as depicted in Figure 4.11 below³.

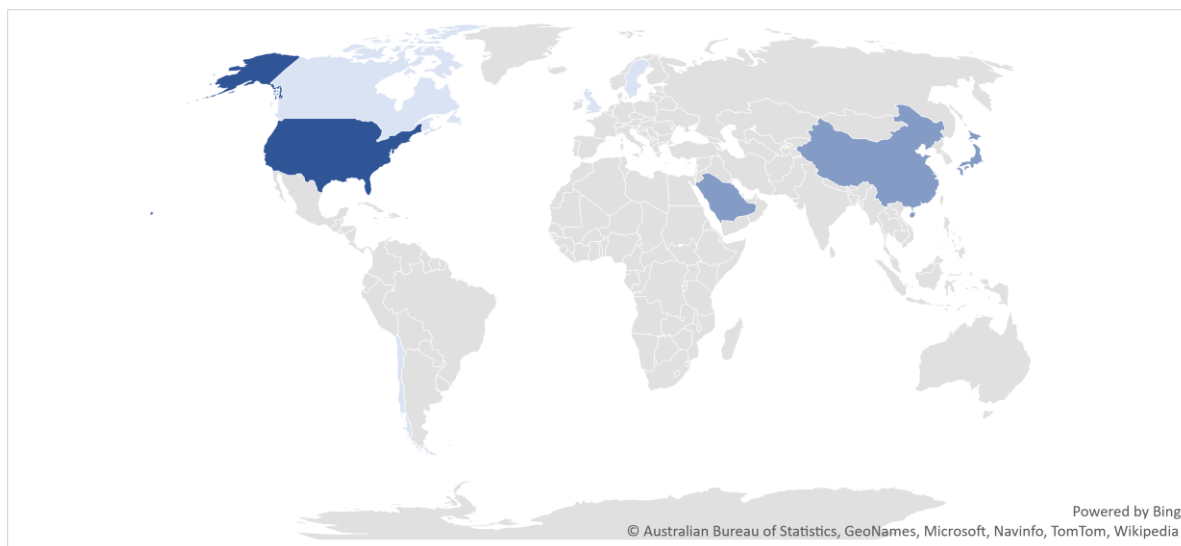


Figure 4.11: Geographical spread of multiple meaning recognition studies

Moreover, similar to research on form-meaning link knowledge Asian countries, such as China (e.g. Li & Kirby, 2015; Zhang & Wen, 2019) and Japan (e.g. Abou-Khalil et al. 2019; Morimoto & Loewen, 2007) are represented in the field of polysemous word knowledge. Arabic countries, such as the United Arab Emirates (e.g. Alnamer, 2017) or Saudi Arabia (e.g. Alqraini & Paul, 2020; Sonbul, El-Dakhs & Al-Otaibi, 2020), also feature as context. Contrastively, European countries appear to be underrepresented at least in this sample of research studies, with Sweden (Karlsson, 2013) and England (Booton et al., 2021) being the only European countries in the selection of studies reviewed.

Comparing the findings of the studies mentioned, resulted in a distinction between educational level. While research studies conducted in the context of secondary school (e.g. Li & Kirby, 2015; Logan & Kieffer, 2021; Logan, 2018) yielded the most promising results regarding polysemous word knowledge, studies conducted at the level of tertiary education report less positive results (Alnamer 2017; Karlsson, 2013; Sonbul, El-Dakhs & Al-Otaibi, 2020; Zhang & Wen, 2019).

Logan's (2018, p. 35) research study involving 45 6th, 7th and 8th grade L2 speakers from a secondary school in the US, generated the highest percentage of correct answers on the Polysemous word test (Logan & Kieffer, 2017), with students selecting 92.67% of the correct polysemous meanings of the 21 high-frequency words covered ($SD=1.42$). Similar findings

³ A more detailed map can be found in appendix 4.3.

are reported by Li and Kirby (2015) and Logan and Kieffer (2021). Featuring 246 Chinese secondary school students Li & Kirby (2015, p. 624) indicate that their participants were able to correctly answer 86.25% ($SD=4.31$) of their test items, which are based on Qian's (2002) Multiple meaning test. Slightly lower but clearly comparable results were detected by Logan & Kieffer (2021, p. 310), who examined polysemous word knowledge of 7th grade students in a longitudinal study located in the US. Using the Polysemous word Test (Logan & Kieffer, 2017, p. 316) they found that their participants were able to correctly identify 77.14% ($SD=3.4$) of the correct multiple meanings of high-frequency words in 7th grade. This result even increased to a correct answer rate of 82.57% ($SD=2.7$) when their participants reached the 8th grade of secondary school. Söderström, Thorén and Karlsson (2022, p. 105) also report a relatively high correct answer rate of 68%. These findings are in agreement with earlier statements by Sullivan (2006, p. 63), who reports that 80 secondary school students, participating in her research study in Canada, made significant progress concerning their polysemous word knowledge from Grade 2 to Grade 11, [$F(3.64)=75.491$, $p<0.001$]. Similar findings are also reported by Schmitt (1998) in a longitudinal study on polysemous word knowledge.

In comparison studies conducted at the level of tertiary education generally yielded much lower results. Alnamer's (2017, p. 116) intermediate Arabic English major students achieved the lowest percentage of 31.2% correctly identified multiple meanings of the verbs *run*, *open* and *make*, directly followed by the advanced students in his study reaching 48.2%. Congruent findings are reported by Karlsson (2013, p. 94) for 15 tertiary level students in Sweden, achieving only 13.2 points out of 40 points, which equates to 33%. Moreover, nearly identical findings were described by Sonbul, El-Dakhs and Al-Otaibi (2020), researching polysemous word knowledge of 60 female students of translation in Saudi Arabia. They state that their participants display rather poor polysemous word knowledge of the 50 phrasal verbs covered, recognizing only 49.12% of the items correctly (Sonbul, El-Dakhs & Al-Otaibi, 2020, p. 7). Productive ability of polysemous word knowledge was even worse with participants being able to produce only 30.57% of the multiple meanings in question, which is lower than 44.5% reported from a similar study by Garnier and Schmitt (2016, p. 34). Slightly more promising results were yielded by Zhang and Wens' (2019) study of 120 English major students in China. Using an acceptability judgment test covering 100 meanings of 50 phrasal verbs, they note that advanced students were able to judge 67.2% of highly frequent phrasal verbs as acceptable, while students with intermediate English language proficiency achieved a correct answer rate of 51.2% (Zhang & Wen, 2019, p. 1085).

However, results appear to change drastically when considering polysemous knowledge of words less frequent in everyday English. Logan and Kieffer (2021) and Logan (2018) compared participants' polysemous word knowledge of general high-frequency words to their ability to correctly identify multiple meanings of academic words. Both studies (Logan, 2018; Logan & Kieffer, 2021), concluded that participants knowledge of multiple meanings

of words, which are highly salient in academic contexts is worse than their polysemous word knowledge of high-frequency words. While Logan's (2018, p. 35) participants correctly identified 65.14% of the meanings, Logan and Kieffers' (2021, p. 316) secondary school students improved from 47.14% in 7th grade to 59.48% at the end of 8th grade. Nevertheless, these results seem to indicate that word frequency has a considerable impact on participant's polysemous word knowledge.

This has been confirmed by several other studies looking into the effect of word frequency on polysemous word knowledge (Booton et al., 2021; Garnier & Schmitt, 2016; Maby, 2016). Booton et al. (2021, p. 15) researched children's polysemous word recognition at primary school level, which confirmed that frequency of word forms contributed to children's knowledge of multiple meanings. Maby (2016, p. 139) confirms these results by showing that participants were more likely to decide that a polysemous sense was acceptable on an acceptability judgement test, if it was more frequently used in English. On a similar strand Verspoor and Lowie (2003) report that participants were more likely to guess a word's secondary meanings, if they knew the most frequent core meaning of the word. The only contradictory result has been found by Karlsson (2013, p. 79), who found no effect of word frequency on polysemous word knowledge. However, she notes that sense-based frequency, being the frequency with which a particular meaning sense of a word is used in the English language, had an impact on test performance, which has been supported by findings of Sonbul, El-Dakhs and Al-Otaibi (2020) and Zhang and Wen (2019). Using a mixed effects model Sonbul, El-Dakhs and Al-Otaibi (2020, p. 1) found that polysemous word knowledge of phrasal verbs was significantly influenced by sense-based frequency as well as proficiency and particle frequency. Similarly, Zhang and Wen (2019, p. 1090) revealed that their participants performed considerably worse on low-frequency senses of the words tested. In addition, Booton et al. (2021, p. 14) reports that dominance and imageability of word meanings also significantly contributes to children's knowledge of multiple word meanings.

However, the research field has not only examined the influence of language related variables on polysemous word knowledge but has also considered the potential influence of biographical variables on performance. While different biographical variables have been considered (e.g. Booton et al., 2021), the influence of Extramural English involvement on polysemous word knowledge is of particular interest for this study. Interestingly, Sonbul, El-Dakhs and Al-Otaibi (2020, p. 8) and Zhang and Wen (2019, p. 1088) yielded similar results, according to which polysemous word knowledge is impacted by the amount of film/video or TV consumption in English participants reported. Additionally, Zhang & Wen (2019, p. 1088) state that time spent on reading English books has a significant impact on advanced students' word knowledge of high-frequency senses of words.

Regarding the influence of particular teaching interventions on polysemous word knowledge rather few studies have been conducted (Abou-Khalil et al., 2019; Algraini & Paul, 2020; Morimoto & Loewen, 2007). A rather basic study, including ten deaf or hard-hearing pupils, was set up by Algraini and Paul (2020, p. 477-478), who investigated whether a teaching intervention using word webs to introduce polysemous words, improved participants' performance. The treatment group, consisting of eight students, outperformed the control group after being taught 75 multiple meaning words through the use of word webs. A slightly more complex study was conducted by Morimoto and Loewen (2007, p. 358), who examined the influence of image-based versus translation-based instruction on participant's polysemous word knowledge. While their 58 secondary school students, divided into two treatment groups and one control group, displayed no significant differences on the pre-test and on the second posttest, the image-based instruction group outperformed the other two groups on the immediate posttest ($p > 0.001$).

In comparison to the two research studies introduced before, the study by Abou-Khalil et al. (2019) is not interested in explicit teaching methods but rather examines students' behavior in an independent setting. Using an informal language learning environment termed SCROLL, Abou-Khalil et al. (2019, p. 5) gathered information about meaning search habits of students when being involved in independent learning situations. They found that 72.18% of the students searched for meanings which were similar to a meaning a learner had encountered in a past experience. Moreover, 75.63% stated that similarity to vocabulary learned shortly before mattered, while 85.05% seemed to learn the most common meaning of a word in the English language (Abou-Khalil et al., 2019, p. 12).

In addition, to considering various factors, which might potentially impact polysemous word knowledge, several researchers examined the actual influence that polysemous word knowledge might have on other English language skills, revealing a clear influence on reading skills (Li & Kirby, 2015; Logan & Kieffer, 2021; Logan, 2018). Using the Gates-MacGinitie reading test Li & Kirby (2015, p. 624) conclude that polysemy has a medium significant effect on reading ability amongst secondary school students, [$r = 0.45$, $p < 0.001$]. Comparable but more narrow results are reported by both of Logan's studies (Logan, 2018; Logan & Kieffer, 2021). Using regression modelling Logan and Kieffer (2021, p.318) reveal that especially knowledge of academic meanings of words has a significant impact on students' reading comprehension in both grades investigated [$\beta = 0.25$, $p = 0.012$; $\beta = 0.28$, $p = 0.024$].

To summarize in relation to the aims of this study prior studies suggest that polysemous word knowledge in general and in particular knowledge of academic meanings is rather poor at tertiary level. Moreover, knowledge of primary meanings should be higher and should aid knowledge of secondary meanings as well.

4.4 What we already know about collocation recognition knowledge

In contrast to other word knowledge components fewer literature can be found on collocation recognition knowledge at the level of primary education (e.g. Miščin, 2015) as well as secondary level (e.g. Miščin, 2015; Shabani & Rahimy, 2020). Instead, most studies appear to have been conducted at tertiary level, which is reflected in the literature reviewed with 29 out of 31 research articles involving participants enrolled in some form of tertiary education (e.g. Bonk, 1996; Dabbagh & Enayat, 2019; Dentisak, 2019; Wolter & Gyllstad, 2013; Ma, 2015; Staehr, 2009). Hence, no comparison of results at different educational levels will be possible for collocation recognition knowledge. Yet, a thorough literature review covering studies published between 1996 and 2020 will be provided. As visible in Figure 4.12 most studies considered have been conducted more recently, with nearly 50% of the studies published between 2015 and 2020.

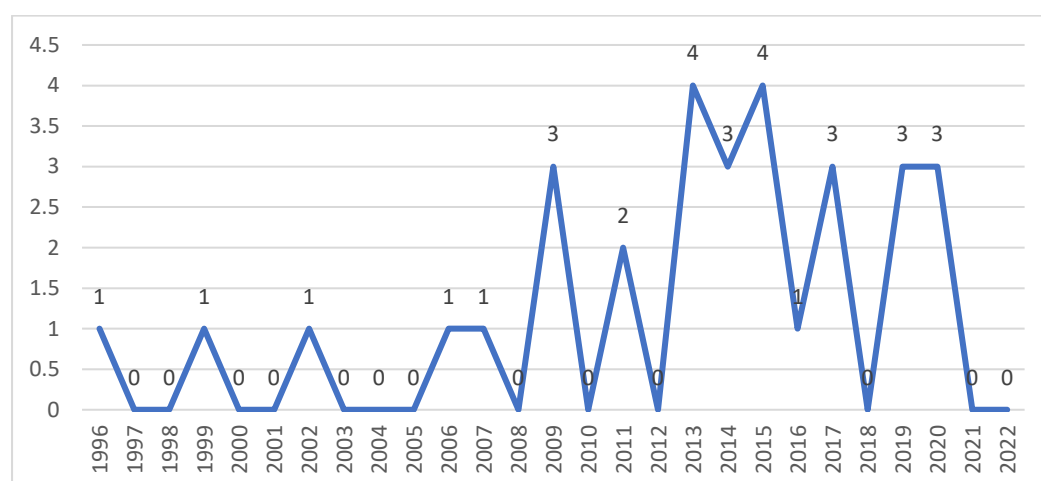


Figure 4.12: Distribution of publication year of collocation recognition studies

Similar to multiple meaning recognition collocation recognition knowledge tends to be measured with the help of self-developed tests as visible in Figure 4.13 below (e.g. Wolter & Gyllstad, 2013; Kamarudin, Abdullah & Abdul Aziz, 2020; Shabani & Rahimy, 2020; Vafaei & Suzuki, 2020). This shows that while for form-meaning recognition knowledge mostly standardized and well-established test formats were used, less well researched word knowledge areas have less widely accepted tests.

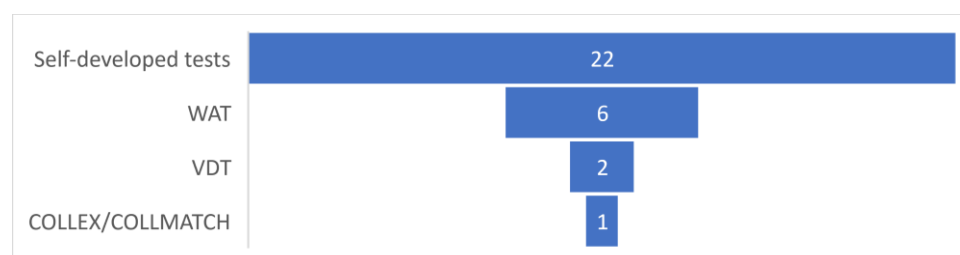


Figure 4.13: Distribution of testing method in collocation recognition studies

However, most self-developed tests follow a similar test design or item type, with some studies using the format of acceptability judgement tests, multiple choice formats or already existent formats like the WAT (Read, 1993) or the Vocabulary Depth Test (Richard, 2011). Overall, older studies appear to rely on published collocation recognition test formats like the WAT (Read, 1993), the Depth of Vocabulary Test (Qian, 2002; Richard, 2011) and COLLEX or COLLMATCH (Gyllstad, 2007). This might suggest that while standard vocabulary tests have been accepted by the wider research community for form-meaning link knowledge, there is still a need for widely applicable collocation recognition tests, which will foster the development of new test formats covering different items and levels of difficulty.

Another difference in comparison to form-meaning link knowledge observable is that target participant groups appear to be even more varied for collocation recognition knowledge than for other word knowledge components researched (see Figure 4.14)⁴. While only one leading country can be identified, which is the Iran with six studies on collocation recognition knowledge amongst tertiary level studies (Adelian, Nemati & Fumani, 2015; Dabbagh & Enayat, 2019; Noori, Gholami & Rajabi, 2014; Shabani & Rahimy, 2020; Talakoob & Koosha, 2017; Vafaei & Suzuki, 2020), the other research studies are spread between a large variety of countries, including European (e.g. Begagić, 2014; Szudarski & Carter, 2016), African (e.g. Nizonkiza, 2015) and Asian countries (e.g. Ma & Lin, 2015; Nguyen & Webb, 2017), as well as parts of the USA.

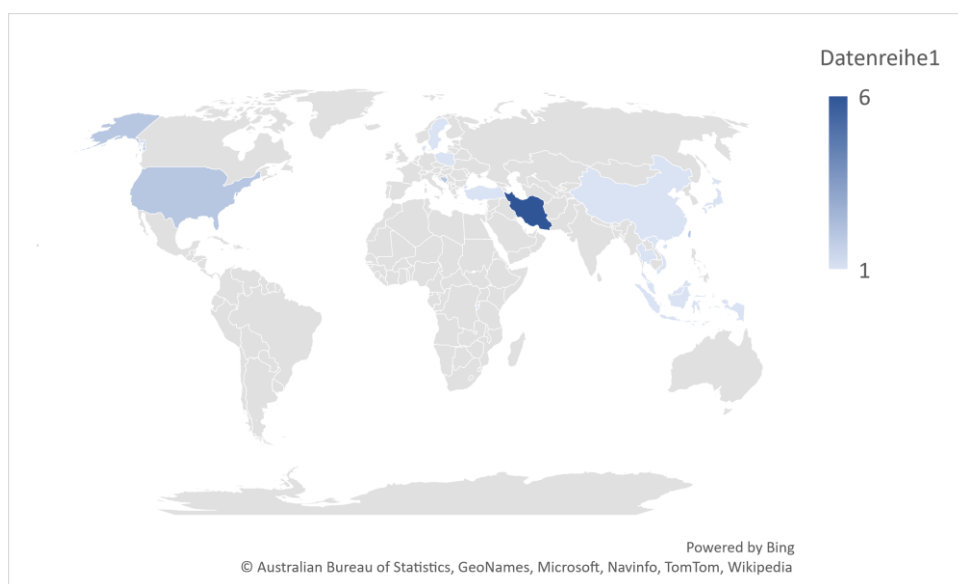


Figure 4.14: Geographical spread of collocation recognition studies

In contrast to Begagić's (2014, p. 46) claim that most studies found rather poor results concerning collocation recognition knowledge amongst tertiary level students, most scores in the literature reviewed appear to range between 40-60% correct answers (e.g. Bonk, 1996; Dabbagh & Enayat, 2019; Dentisak, 2019; Ma, 2015; Staehr, 2009). Vafaei and

⁴ A more detailed map can be found in appendix 4.4.

Suzukis' (2020, p. 398) study of 263 Iranian tertiary level students using an aural version of the VDT resulted in participants being able to give 48.96% of the answers correctly. Very similar results are also reported by Dentisak (2019, p. 118), who found that 300 Thai students pursuing tertiary education were able to identify 42.47% of the items in his self-developed acceptability judgement test correctly. Additionally, Nguyen and Webb (2017, p. 311) and Nurweni and Read (1999, p. 166) both using multiple choice test formats found comparable results with 45.72% and 45% correct answers. While a correct answer rate of below 50% could indeed be interpreted as a rather low result in comparison to other word knowledge components, only Webb and Kagimoto's (2009, p. 64) research study including 145 university students in Japan appears to really validate Begagić's (2014) view with a correct answer percentage of only 17.71% on the pre-test. However, also in Webb and Kagimotos' (2009, p. 64) study students achieved higher scores on the post-test. Overall, collocation recognition test results tend to even increase when looking at the literature, with many studies stating that their participants were able to answer between 60-80% of all items correctly (e.g. Nizonkiza, 2015; Kamarudin, Abdullah & Abdul Aziz, 2020; Talakoob & Koosha, 2017).

The highest result achieved is reported by Kamarudin, Abdullah and Abdul Aziz (2020, p. 3) for 21 Malaysian university students taking a self-developed receptive and productive collocation test. Out of 48 target collocations participants were able to correctly identify 46.14, which equates 96.13% (Kamarudin, Abdullah & Abdul Aziz, 2020, p. 4). Comparable results were only found by Gyllstad (2007, p. 208) who reports that 24 upper secondary school students achieved 84% correct answers on the COLLEX and 80% on the COLLMATCH. While the tests of both studies were carefully developed they worked with small numbers of participants. In contrast other studies such as Qian (2002, p. 522-523) including 217 Arabic speaking students or Wang (2014, p. 2460) investigating collocation recognition knowledge of 168 Chinese students have built their studies on more solid ground. Nevertheless, both studies reveal slightly lower but comparable results with 70.66% correct answers on the DVK (Qian, 2002, p. 528) and 67.5% correct answers on the WAT (Wang, 2014, p. 2461). Other more recent studies conducted in Iran also yield comparable findings with 72.06% (Talakoob & Koosha, 2017, p. 14) and 70.7% (Adelian, Nemati & Fumani, 2015, p. 977) correctly identified answers on the self-developed collocation recognition tests.

While most studies cover several different types of collocations some studies pay particular attention to knowledge of different collocation types (e.g. Alsakran, 2011; Barfield, 2006; Begagić, 2014; Dentisak, 2019; Kamarudin, Abdullah & Abdul Aziz, 2020). Li and Schmitt (2009) focus on the narrower scope of academic lexical phrases in a case study of one Chinese MA student. By analysing her texts written over the timespan of one year additionally to interviews about each text written they find that their participant displays considerable improvement of the use of academic lexical phrases, learning 166 new formulaic sequences (Li & Schmitt, 2009, p. 85). However, as often observed in L2 learners

the participant also displays an overuse of familiar and in her eyes secure academic lexical phrases (Li & Schmitt, 2009, p. 85).

Focusing on general collocations Kamarudin, Abdulla and Abdul Aziz (2020, p. 4) investigated three types of collocations, namely verb + noun, verb + preposition and adjective + noun, finding that verb + noun collocations were answered correctly most often, followed by adjective + noun collocations and verb + preposition collocations being the hardest for their participants. Similar findings were detected by Alsakran (2011, p. 62) and Dentisak (2019, p. 119), with students answering verb + noun collocations correctly most often. While Alsakran only compares three types of collocation (verb + noun, adjective + noun and verb + preposition) Dentisak (2019) examining differences in knowledge between grammatical (noun + preposition, adjective + preposition, verb + preposition and other verb structures) and lexical collocations (verb + noun and adjective + noun) resulting in a similar preference for verb + noun collocations. Therefore, since this study exclusively covers verb + noun collocations, higher results than for some of the studies including also other collocation types might be expected.

Similarly, effects regarding the frequency of the node word might be expected, since some studies, such as Nguyen and Webb (2017, p. 298) and Nizonkiza (2015, p. 125) found significant differences in performance regarding word frequency of the node word. However, a study by Wolter and Gyllstad (2013, p. 464) of congruent collocations, having an L1 correspondance and incongruent collocations, having no L1 corespondance, shows that instead of word frequency collocational frequency has a significant effect on students' results. Yet, in contrast to other studies Wolter and Gyllstad (2013, p. 460) used an acceptability judgement task, asking the learners to decide whether items are commonly use in English or not and were mainly interested in students' response time and error rate. Hence, while prior studies focused on the number of correct responses Wolter and Gyllstad (2013) work with the assumption that the quicker students respond to an item, the easier the task was for them. By comparing 25 native English speakers against 25 non-native speakers (L1 Swedish) of English they reveal that native speakers only significantly ($p=0.003$) outperformed non-native speakers regarding incongruent collocations (Wolter & Gyllstad, 2013, p. 463). Furthermore, non-native speakers responded to congruent collocations significantly ($p=0.046$) faster than to incongruent collocations. In a related study Gyllstad and Wolter (2016, p. 304) investigate the reaction times of 27 advanced Swedish L2 English learners when judging collocations and free combinations in comparison to 28 English native speakers. Comparing the reaction times revealed that both groups needed more time to react to collocations in comparison to free combinations (Gyllstad & Wolter, 2016, p. 304), which was attributed to two factors. Firstly, participants might need more time to react to collocations, due to their "semantically semi-transparent nature" (Gyllstad & Wolter, 2016, p. 296). Moreover, phrasal frequency appear to predict reaction times as well.

In addition to studies using collocation tests recent studies start using eye-tracking technology to reach conclusions concerning collocational knowledge. In their study of forty advanced non natives speakers of English Vilkaitė and Schmitt (2019, p. 335) analyse their participants eye movements when reading adjacent and non-adjacent collocations. While adjacent collocations show a significant processing advantage over matched control phrases concerning both total reading time and fixation, non-adjacent collocations display a similar trend but without reaching significance (Vilkaitė & Schmitt, 2019, p. 345). Hence, non-adjacent collocations seem to be processed faster than control phrases but not to the same extent as adjacent collocations.

Moreover, it was found that overall language proficiency significantly influences collocation knowledge (e.g. Nizonkiza, 2015; Talakoob & Koosha, 2017; Webb & Kagimoto, 2009). Nizonkiza (2015, p. 135), including 115 tertiary level students from Burundi, has found that the results on the TOEFL test, indicating overall language proficiency was significantly related to participants' performance on the collocation test. Moreover, in a study of the effect of receptive and productive learning tasks on collocational knowledge, Webb and Kagimoto (2009, p. 360) found that for more proficient learners productive task types appeared to be more beneficial, while for less proficient language learners receptive tasks lead to more success. In a similar study of 134 undergraduate students in Turkey Özdem Ertürk (2017, p. 67) reports that both receptive and productive learning tasks lead to more progress than in the control group, who got no input. This is confirmed by Shabani and Rahimy (2020, p. 1019) who also found that input through textual glossing, audio glossing or skewed input led to more success than the control group receiving no special treatment. In addition, in Özdem Ertürk's (2017, p. 67) population of student's receptive tasks ($M=12.52$) led to significantly ($p<0.001$) better results on the multiple-choice collocation recognition test than productive tasks ($M=10.45$).

Apart from intentional attempts at learning Noori, Gholami and Rajabi (2014) and Webb, Newton and Chang (2013) investigated whether incidental learning can have an effect on receptive collocation knowledge. Noori, Gholami and Rajabi (2014, p. 148) investigated 50 tertiary students in Iran, who were split into two groups, the first group receiving explicit teaching and the second group receiving incidental input to collocations. They found that the students who encountered collocations implicitly achieved better scores on the collocation recognition test (Noori, Gholami & Rajabi, 2014, p. 150-151). Similarly, Webb, Newton and Chang (2013, p. 108) found significant improvement of collocation recognition scores after their 161 participants read texts including the collocations covered in the test. Consequently, they claim that "if an approach were taken to include useful collocations in graded reading schemes, there may be little need to teach collocations explicitly for learners actively taking part in extensive reading programs" (Webb, Newton & Chang, 2013, p. 111). Nonetheless, a study by Sonbul and Schmitt (2013) shows that also explicit instruction can lead to learning gains regarding collocational knowledge. They investigated implicit and

explicit collocational knowledge of 43 postgraduate students and whether these types of knowledge can be enhanced by either enriched (collocations in a text), enhanced (highlighted collocations in a text) or decontextualized treatments (collocations presented in isolation) (Sonbul & Schmitt, 2013, p. 143). Their analysis showed that all conditions lead to an improvement of explicit collocational knowledge. However, none of the conditions seemed to influence implicit collocational knowledge (Sonbul & Schmitt, 2013, p. 150).

Linking the results of this literature review to this study implies that while most participants in the studies reviewed achieve a correct answer rate between 40%-60%, the choice of exclusively verb + noun collocations, as well as the overall high proficiency level of participants might lead to comparably high results. Moreover, both implicit and explicit learning are able to positively enhance collocation recognition knowledge.

4.5 Potential relationships between word knowledge components researched

After reviewing each word knowledge component considered in this study individually, this chapter will attempt to provide a summary of relationships reported between the four word knowledge aspects. González-Fernández and Schmitt (2019) attempted to examine the relationship between eight word knowledge components, namely form-meaning recognition and recall knowledge, collocation recognition and recall knowledge, multiple meaning recognition and recall knowledge and derivative recognition and recall knowledge. By using structural equation modeling they found that the recognition and recall knowledge of each word knowledge component should be rather viewed as an individual component adding to overall vocabulary knowledge, with recall knowledge tending to contribute slightly more than recognition knowledge (González-Fernández & Schmitt, 2019, p. 15). However, this is not true for form-meaning knowledge, in which recognition knowledge contributes slightly more to the overall dimension of vocabulary knowledge than recall knowledge as can be seen in Figure 4.15 below. In general, “the factor loadings of this model are very 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” (González-Fernández & Schmitt, 2019, p. 16).

In an extended replication study of this study, González-Fernández (2021) investigates the same four word knowledge components both receptively and productively. Taking a look at the participants’ correct answer rate for the individual components reveals that form-meaning recognition knowledge is the best known component with 80.96% correct answers, followed by collocation recognition knowledge (78.41%), multiple meaning recognition knowledge (67.60%), derivative recognition knowledge (63.65%), collocation recall knowledge (58.79%), form-meaning recall knowledge (53.36%), derivatives recall knowledge (51.60%) and lastly multiple meaning recall knowledge (50.01%) (González-Fernández, 2021,

p. 34). Therefore, the results of this study align with the prior finding that recognition knowledge is larger than recall knowledge. Including 314 EFL learners with L1 Spanish and L1 Chinese in her study she reports that when using a Friedman test analysis and a post hoc Wilcoxon signed-rank test with Bonferroni correction all word knowledge components were significantly different from each other ($p < 0.001$) (González-Fernández, 2021, p. 34). Hence, while not using the same statistical means as in González-Fernández and Schmitt (2019) these results might suggest that in this wide population word knowledge components appear separately.

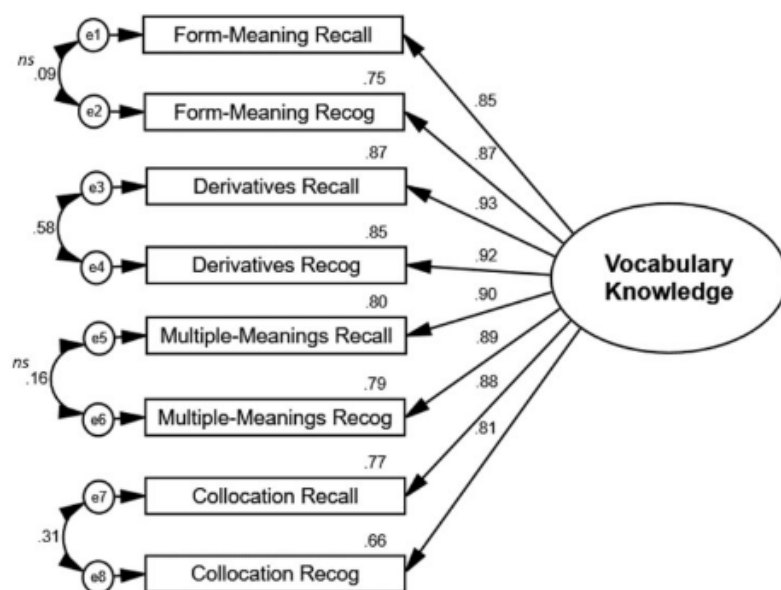


Figure 4.15: SEM model with best fit indices by González-Fernández & Schmitt (2019, p. 16)

Since no other studies apart from González-Fernández and Schmitt (2019) and González-Fernández (2021) investigated more than two word knowledge components at a time all other findings will be reported for the relationships between two word knowledge components below. As visible in Table 4.2 below, most findings were reported regarding the connection between form-meaning recognition and collocation recognition, spanning a relatively large range of correlation coefficients.

Number of studies	Word knowledge aspects researched	Maximum	Minimum
8	Form-meaning recognition & Form-meaning recall	$r = 0.84$ (Zhang & Lu, 2015)	$r = 0.41$ (Nizonkiza, 2015)
5	Form-meaning recognition & Multiple meaning recognition	$r = 0.67$ (Lawrence et. al, 2019)	$r = 0.35$ (Li & Kirby, 2015)
12	Form-meaning recognition & Collocation recognition	$r = 0.93$ (Ehsanzahdeh, 2016)	$r = 0.31$ (Zhong, 2018)
2	Form-meaning recall & Collocation recognition	$r = 0.55$ (Zhang & Lu, 2015)	$r = 0.46$ (Zhong, 2018)

Table 4.2: Relationships between word knowledge components reported in the literature

The strongest relationship between form-meaning recognition and collocation recognition was identified by Ehsanzahdeh (2016, p. 32) in her study of 33 undergraduate students in Iran, using two well established vocabulary tests, namely the VLT (Schmitt, Schmitt & Clapham, 2001) and the WAT (Read, 2004). While Ehsanzahdeh's (2016) relatively low participant number might leave some room for doubt, a comparable result, with $r = 0.87$, $p < 0.001$, was achieved by Vafaei & Suzuki (2020, p. 399), who examined receptive vocabulary size and receptive collocation knowledge of 263 Iranian tertiary level students, using a similar methodological approach. Moreover, Nurweni and Read (1999) and Staehr (2009) report similarly large correlation coefficients with $r = 0.79$, $p < 0.001$ (Nurweni & Read, 1999, p. 167), and $r = 0.80$, $p < 0.001$ (Staehr, 2009, p. 591) in the context of Indonesia and Denmark, indicating that form-meaning recognition knowledge and collocation recognition knowledge appear to be strongly related. Interestingly, all four studies mentioned have used the same or very comparable measurement instruments, namely the VLT for receptive vocabulary size and the WAT (Ehsanzahdeh, 2016; Nurweni & Read, 1999; Vafaei & Suzuki, 2020) or the Depth of Vocabulary Knowledge Test (Staehr, 2009), which has been developed on the basis of the WAT format.

Slightly lower correlation coefficients were observed by Qian (2002), Nguyen and Webb (2017), Wang (2014) and Lee (2020) ranging between $r = 0.68$, $p < 0.01$ (Qian, 2002, p. 529) and $r = 0.61$, being part of Lee's (2020, p. 8) SEM model. In comparison to the studies mentioned above these studies used a methodological mix, with Nguyen and Webb (2017) designing a self-developed collocation test, Qian (2002) using the DVK, Wang (2014) mixing old and new WAT items and Lee (2020) utilizing the classical WAT version. Nevertheless, measurement instruments applied are still very comparable, since the WAT and adapted version, such as the DVK used by Vafaei & Suzuki (2020), Lee (2020) Ehsanzahdeh (2016), Wang (2014), Staehr (2009), Qian (2002) and Nurweni & Read (1999) cover highly frequent general and academic adjective + noun collocations (Qian, 2002; Read, 1993). Only Nguyen and Webb's (2017) self-developed multiple-choice test covering 180 items is concerned with verb + noun collocations, similar to the Academic Collocation Test used in this study.

Slightly smaller but still large correlation coefficients higher than $r = 0.50$ were identified by Ma and Lin (2015, p. 8) [$r = 0.52$, $p < 0.001$] and Enayat & Amirian (2020). In addition, Enayat and Amirians' (2020, p. 105) study revealed that different proficiency levels, ranging from low-intermediate to advanced English language ability, considerably differed regarding strength of relationship between the two word knowledge aspects researched. While for low-intermediate students form-meaning recognition knowledge was still strongly correlated with their collocation recognition knowledge with medium effect size [$r = 0.51$, $p < 0.001$, $r^2 = 0.26$], upper-intermediate students' word knowledge components only displayed a medium correlation [$r = 0.40$, $p < 0.001$, $r^2 = 0.16$] with small effect size, while advanced students displayed no significant relationship at all [$r = 0.345$, $p = 0.115$, $r^2 = 0.12$]

(Enayat & Amirian, 2020, p. 105). This implies that overall language proficiency level might have a significant impact on the strength of relationships between form-meaning recognition and collocation recognition knowledge.

Out of all studies reviewed eight studies investigated the relationship between form-meaning recognition and form-meaning recall knowledge. Out of these, three studies (Schwarz, 2020; Zhong, 2018; Zichtl, 2017) were conducted with secondary school students and four studies (Cheng & Matthews, 2018; Miralpeix & Muñoz, 2018; Nizonkiza, 2016; Zhang & Lu, 2015) involved tertiary level participants. In comparison the results for secondary school students span a smaller range of results with correlation coefficients lying between $r = 0.64$, $p < 0.001$ (Zhong, 2018, p. 365) and $r = 0.53$, $p < 0.01$ (Schwarz, 2020, p. 244) then findings for tertiary school students. Regarding the tertiary level the strongest correlation was found by Zhang and Lu (2015, p. 748) with $r = 0.84$, $p < 0.01$, for 150 first year students in China, who additionally used SEM modelling in their study. By contrast the smallest result reported by Nizonkiza (2016, p. 178) is less than half of Zhang and Lu's (2015) finding. Nizonkiza (2016, p. 178) states that the relationship between receptive and productive form-meaning knowledge is of medium strength with $[r = 0.40, p < 0.01]$ for 115 tertiary level students in Burundi. According to these findings the relationship between form meaning recognition and form-meaning recall is more diverse in the population of tertiary level students, with some students showing a very strong tie between the two word knowledge components and others displaying a similar strength of relationship as found for secondary school students. This might imply that the strength of the link between form meaning recognition and form meaning recall knowledge at tertiary level depends on particular predictor variables, such as type of instruction or vocabulary input, which might influence the connectedness of the two word knowledge components.

The third most researched connection between the word knowledge aspects under investigation is the relationship between form-meaning recognition and multiple meaning recognition knowledge with a total of five studies dealing with these components. In a large scale study featuring 5855 secondary school students in the US Lawrence et al. (2019, p. 296) detected the strongest relationship between the two word knowledge components $[r = 0.67, p < 0.01]$. However, while Logan and Kieffer's (2021) and Li & Kirby's (2015) research on secondary school level, yielding considerably smaller results with $r = 0.47$, $p < 0.01$ (Logan and Kieffer, 2021, p. 317) and $r = 0.35$, $p < 0.01$ (Li & Kirby, 2015, p. 624), focus on general receptive form-meaning and polysemous word knowledge, Lawrence et al. (2019) place a particular focus on academic vocabulary. Hence, the remarkable difference between these findings might be explained through the diverse focus on different vocabulary items.

Taking a look at results from the tertiary level reveals that at this stage of educational development participants depict a similar relationship between the two word knowledge components as reported by Lawrence et al. (2019). Investigating 107 college students'

general knowledge of form-meaning recognition and polysemous word knowledge, Binder et al. (2017, p. 337) conclude that the two word knowledge components are strongly connected [$r=0.63$, $p<0.01$]. While based on a rather thin basis of research studies, these findings might infer that the link between form-meaning recognition knowledge and multiple meaning recognition develops over time.

Lastly, only two studies reviewed (Zhang & Lu, 2015; Zhong, 2018) present findings regarding the relationship between form-meaning recall and collocation recognition knowledge, both resulting in comparable findings. Similar to Zhang and Lu (2015, p. 748), who note that the strength between the two components was of medium strength ($r=0.55$, $p<0.01$), Zhong (2018, p. 365) observes a medium correlation of $r=0.46$, $p<0.01$ between the two word knowledge aspects. While there is already considerably less literature dealing with form-meaning recall knowledge and collocation recognition knowledge, even less research can be found on the connection between form-meaning recall and multiple meaning recognition knowledge. This emphasizes the need for more examination of form-meaning recall knowledge and its potential links to other word knowledge components in order to balance the dominance of form-meaning recognition knowledge in the research field.

5 Language learning beyond the classroom

While research on L2 learning has long focused on classroom settings, the rise of new media has triggered increased interest in language learning outside of an institutional setting (Sylvén, 2019a, p. 152). Although young children mostly name spare time activities such as playing outdoors and meeting friends as their favourite spare time activities in both Germany and Austria, they already report an increasing interest in watching television (Feierabend, Plankenhorn & Rathgeb, 2014; education group, 2018), indicating a significant trend towards the use of digital devices in the German-speaking area. According to surveys by the Medienpädagogischer Forschungsverbund Südwest [media-pedagogic research collective South-West] (2014) and by the education group (2018), for children aged 2-3, books are the most indispensable medium in Germany, whereas children aged 4-10 do not want to give up watching television in both Germany and Austria. For older children aged between 11 and 19, a slightly different trend can be detected. While social contacts are still crucial for young people in Germany and Austria, they report less interest in watching TV (Medienpädagogischer Forschungsverbund Südwest, 2019; education group, 2019). Instead smartphones gain importance as a technical device that can be used for several activities, such as communication, playing games, or watching videos. Consequently, smartphones are reported to be the device that young adults do not want to do without. Another visible trend is that young people steadily have more unlimited internet access, which is mostly used via their smartphone (Medienpädagogischer Forschungsverbund Südwest, 2019; education group, 2019). In Germany, young people reported that they spent about 205 minutes per day online (Medienpädagogischer Forschungsverbund Südwest, 2019).

Regarding the activities done online, a shift from searching for information and communication to playing games and entertainment has crystalized (Medienpädagogischer Forschungsverbund Südwest, 2019). In Austria online consumption of audio-visual input has doubled from 1.5h to 3h (education group, 2019). However, girls tend to use the internet more for the purpose of communication and spend less time playing computer games compared to boys their age (Medienpädagogischer Forschungsverbund Südwest, 2019; education group, 2019).

Reacting to students' increased participation in digital spare time activities, more researchers have started to investigate their potential as new sources for language learning aside traditional media (Sundqvist & Sylvén, 2014). However, the increasing number of scholars involved in the field does not only lead to more findings but, additionally, to an abundance of labels and different use and understanding of terminology. Next to *Extramural English* (EE), one of the theoretical concepts investigated in the present study (see chapter 5.1), other terms such as *out-of-class learning* (Benson, 2001; Cole & Vanderplank, 2016; Verspoor, De Bot & Van Rein, 2011; Yap, 1998), *after-class learning*, *extracurricular learning* (Reinders & Benson, 2017), *out-of-school learning* (Lamb, 2004b; Yi, 2005), *informal*, *non-formal* and *formal learning* (Schugurensky, 2000), *naturalistic learning*, *independent learning*, *uninstructed learning* (Housen & Pierrard, 2005), *implicit learning* (DeKeyser, 2003; Ellis, 2009), *incidental learning* (Laufer & Hulstijn, 2001), *unintentional learning* (Forsman, 2004) or *spare-time English* (Lundahl, 2009), amongst others, can be found. While some of these terms can be understood literally, often referring to synonymous concepts, other terms are based on complex theory.

Looking at the perspective of the concepts mentioned, many view language learning from the angle of the physical or temporal setting of the learning activity (e.g. *after-class learning*, *spare time English*, *extracurricular learning*, *out-of-school learning*). Other concepts, in contrast, reflect on language learning from the perspective of learner awareness. For example, *incidental language learning* refers to “learning without an intent to learn” (Laufer & Hulstijn, 2001, p.10), as does the concept of *implicit language learning*, which has been described as learning “without either intentionality or awareness” (Ellis, 2009, p. 7). As a third example, Forsman’s (2004, p.173) term *unintentional learning*, being “accidental learning of information without the intention of remembering that information”, seems to take a similar stance. Considering the roles of intentional and accidental learning, Schugurensky (2000) introduces his concept of *informal learning*. As can be seen in Figure 5.1, he conceptualises informal learning as comprising various degrees of conscious and intentional learning experiences.

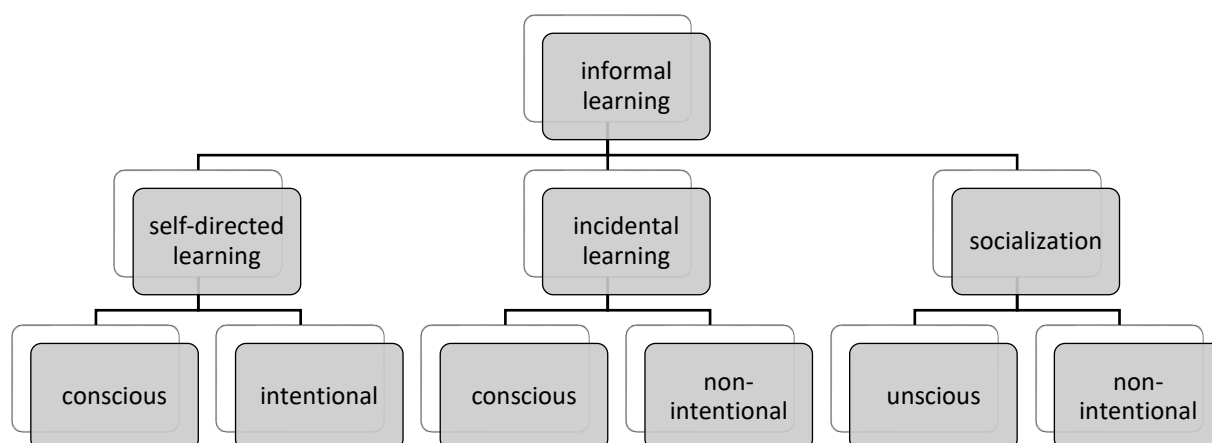


Figure 5.1: Informal learning according to Schugurensky (2000)

He states that *informal learning* can consist of either *self-directed learning*, being both conscious and intentional on the part of the learner, *incidental learning*, i.e., conscious but not intentional, or *socialization*, which is both unconscious and non-intentional. While Schugurensky's (2000) model is a useful conceptualisation of different levels of awareness involved in learning processes outside a classroom setting, several important variables seem to be missing. Providing a more complete picture of the learning process both in- and outside of a traditional school setting, Benson (2011) has developed a multifaceted model on *language learning and teaching beyond the classroom* (LBC), which builds on his own definition of *out-of-class learning* as "any kind of learning that takes place outside the classroom and involves self-instruction, naturalistic learning or self-directed naturalistic learning" (Benson, 2001, p. 62) In this model, Benson (2011, p. 9) differentiates four main dimensions of learning within and beyond the classroom, namely *location*, *formality*, *pedagogy* and *locus of control*, which has been visualized by Schwarz (2020) as can be seen in Figure 5.2 below.

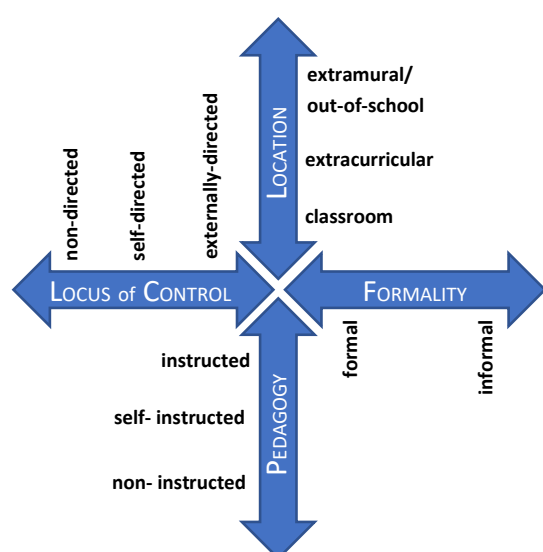


Figure 5.2: Adapted LBC model of Benson (2011) and Schwarz (2020)

The concept of *location* considers the phenomenon of learning in different settings from a spatial point of view, with one end of the continuum being the classroom in a traditional school setting and the other pole being extramural or out-of-school environments. Extracurricular activities still taking place at school but outside of teaching time are situated in between the two extremes. However, it is crucial to note that, as Schwarz (2020, p. 8) highlights, Benson (2011) originally “uses the term extramural as a synonym for extracurricular”. This differentiates his conceptualisation of this term considerably from Sundqvist’s (2009) concept of *Extramural English*, used in the present study. To be more precise, Sundqvist describes extramural activities as being “exposure to the target language outside of the educational context” (Sylvén, 2019a, p. 152), which is why Schwarz (2020) has moved the term to the far end of the continuum, synonymously to out-of-school learning. This equation appears to be quite suitable since Sundqvist and Sylvén (2016, p. 8) state that Benson’s (2011) concept of *out-of-school learning* is comparable to *extramural English*.

The next dimension of Benson’s (2011) model, *formality*, refers to “the degree to which learning is linked to educational qualifications or structured by educational institutions” (Reinders & Benson, 2017, p. 561). While formal learning situations are directly relevant for assessment purposes and consequently for the achievement of qualifications, informal learning situations appear to have no immediate impact on qualifications. Additionally to the two poles, *formal* and *informal learning*, he places non-formal learning in the middle of the continuum, being “classroom or school-based programmes that are taken for interest and do not involve tests or qualifications” (Benson, 2011, p. 10). However, similar to Schwarz (2020), this study will not include the concept of non-formal learning, since the two terms, informal and non-formal learning, are frequently used synonymously in practice. Instead this study will adopt Schwarz’s (2020, p. 9) suggestion to view the dimension of formality as a continuum.

The third dimension of the model, namely *pedagogy*, introduces another continuum ranging from *instructed*, over *self-instructed* to *non-instructed* settings. While Benson (2011, p. 11) states that instructed settings involve “formal processes such as sequencing of material, explicit explanation and testing”, this study argues that instruction should be only defined by the amount of teaching taking place, leaving aside the aspect of assessment, which is already included in the dimension of formality. Hence, instructed learning involves a large amount of teaching by a person other than the learner. In situations in which learners largely instruct themselves, such as in exploratory tasks, one can speak of self-instructed settings. Non-instructed settings of learning are unguided and thus entail “no intention to learn” (Benson, 2011, p. 11). Therefore, the dimension of pedagogy is not exactly a continuum of the amount of teaching input given but rather of the agency of learning, ranging from externally instructed over self-instructed to non-instructed learning processes.

The last dimension included in Benson's (2011) model of learning beyond the classroom is *locus of control*. Schwarz (2020, p. 10) describes this concept as "who holds control over learning", ranging from teacher-directed over self-directed to non-directed learning. Using the example of a tutoring lesson, she explains that this learning situation would be extramural, rather formal, instructed and teacher-directed, with the teacher holding "most control over the decisions for learning" (Schwarz, 2020, p. 10). However, following this definition, *locus of control* appears to be overlapping with the dimension of *pedagogy*, which is concerned with the agency in learning processes. Hence, if teacher-directed learning means that the teachers decides how learning takes places, this would entail instructed learning, where the teacher acts as an agent giving instructions. For instance, if the tutor asks his/her students to look at three existent book reports to explore the structure of this text type autonomously during the tutoring lesson, this learning situation could be described self-instructed, since the students teach themselves by having a look at the texts. Hence, the students decide how learning takes place in this situation, which would make the situation a self-directed learning situation. Nevertheless, one could argue that the students do not complete the exercise for themselves but for the teacher, who controls the learning situation. Conversely, if a student looks up an unknown word in a dictionary self-instructedly to learn the word, the learning process is not only initiated by the person him/herself but also completed for him/herself. Consequently, this study opts for a new definition of the dimension *locus of control*. Instead of defining the dimension by who holds control over the learning situation, the concept should be more explicitly guided by the agent the learning is directed at. In other, simpler words: Who does the learner learn for? Moreover, it is proposed that the term *teacher-directed* is changed to *externally-directed*, since learning cannot only be directed at teachers but also by other agents, such as parents, friends or partners.

While building his model on the four dimensions mentioned, Benson (2011, p. 13) advises researchers to pay additional attention to the social and pedagogical conditions of learning situations and the availability and use of resources for learning. Reinders and Benson (2017, p. 563) add that other important dimensions to be considered are *sociality*, *modality*, *mediation*, and *linguistics*. While the dimension of *sociality* covers "the social relationships and networks involved in the learning processes" (Reinders & Benson, 2017, p. 563), *modality* is concerned with "the learning practices [students are] engaged in" (Reinders & Benson, 2017, p. 563). The dimension of *mediation* covers the materials used in the learning activity. Moreover, the concept of *linguistics* should consider the linguistic preconditions of the learner, such as their proficiency level (Reinders & Benson, 2017, p. 563). While *sociality* as defined by Reinders and Benson (2017) appears to be too complex to integrate into a model, a dimension which gives information about the extent of social contact involved in a learning activity would be beneficial, since this category would allow at least a rough estimation of how many learning activities are carried out alone, in pairs or in groups.

Additionally, a dimension regarding the learning material or resources used could be implemented, similar to Reinders and Bensons' (2017) concept of *mediation*. This would allow researchers to detect tendencies where and for which learning activities which medium is used. A suggestion for such a (preliminary) model of language learning in and beyond the classroom, which incorporates the named dimensions, is provided below (see Figure 5.3).

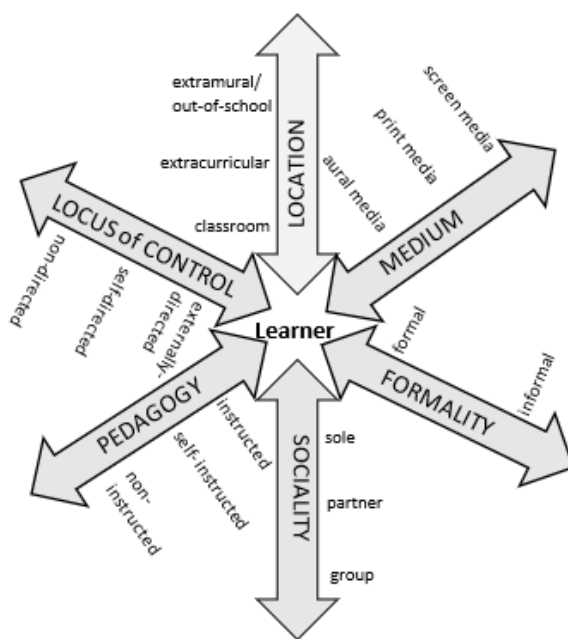
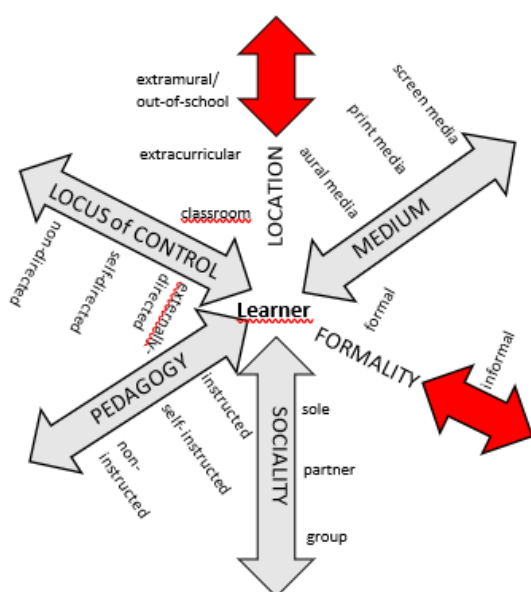


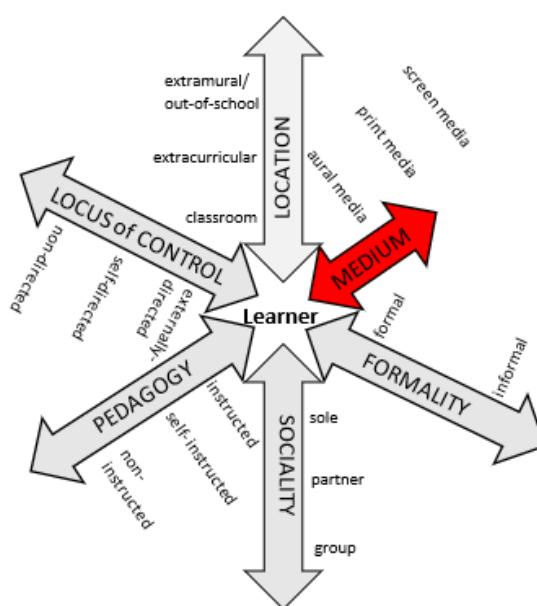
Figure 5.3: Model of language learning in and beyond the classroom

Based on Hartley's (2020, p. 201) definition of the words *media* and *medium*, the dimension of *medium* added to this model covers aural media, being recorded music or speech, print media, such as newspapers or books, and screen media, including multimedia input such as watching television or playing computer games. In addition to the two added dimensions, the preliminary model puts the learner with all her/his preconditions in the centre. While some learners will display the whole range of each dimension, others might be limited by external factors. For instance, for a pre-school child mostly extramural, rather informal activities will be possible, reducing these dimensions considerably as visualized by the red arrows in example 21. Another example might be a language learner with no internet access, who will only be able to use aural or print resources (see example 22).

Example 21



Example 22



Additionally, this model, similar to Benson's (2011) original model, allows researchers to describe and assign learning activities. This might, for example, be helpful for creating learner profiles to examine whether different learners display different learning behaviours. To exemplify this, Figure 5.4 presents the fictional learning profile of Hannah. Blue lines represent her extramural activities, while red lines indicate her in-class activities. The line width shows how often a particular learning behaviour was reported.

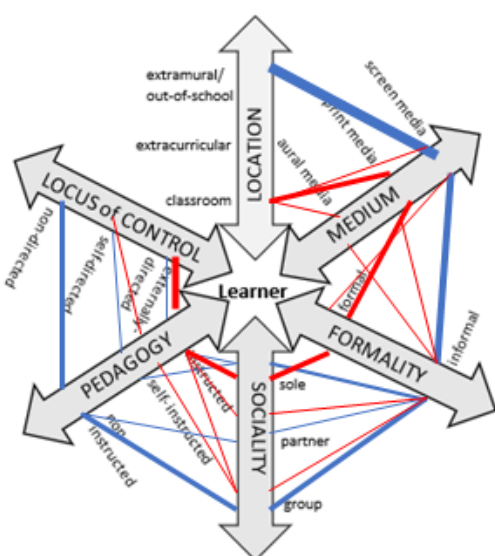


Figure 5.4: Using the model of language learning in and beyond the classroom for learning profiles (e.g., Hannah)

The example in Figure 5.4 shows that Hannah is solely using digital media for her extramural activities. Most of these extramural activities tend to be quite informal and not related to any form of educational assessment. While most of the extramural English activities are carried out digitally in groups (e.g. when playing multiplayer online games or chatting on

Facebook), some tasks are carried out alone (e.g. watching music videos or reading newspaper articles in the New York Times App), and very few tasks are completed in pair work (e.g., playing SIMS and chatting). While all group and partner activities of Hannah are non-instructed without a conscious intention to learn, individual activities are both self-instructed (e.g., consciously recording new words when reading online) or instructed (e.g., recording vocabulary for a reading portfolio for school). All non-instructed activities are also non-directed, which means that no controlled, conscious learning takes place but rather what Schugurensky (2000) calls *socialization*, i.e., non-intentional and unconscious language learning. Self-instructed learning tends to be self-directed and instructed learning tends to be externally directed.

At school, Hannah uses mostly print media. However, digital media or aural media are seldom incorporated. Activities utilizing digital media are solely formal (e.g. check-ups on Quizlet or PowerPoint slides used for in-class presentations), similar to activities using print media, which are mostly formal (e.g., exercise in the textbook, reading comprehensions). Activities using aural media are exclusively informal (e.g., discussions in class). To summarize, while informal activities are tendentially performed in groups or with a partner, formal activities in school are carried out alone. Only group activities are partially self-instructed and self-directed; all other activities are instructed and externally-directed.

While the model is believed to aid researchers in their process of recording, analysing and interpreting learning activities in and beyond the classroom, one must not forget that all models are simplifications of complex phenomena, such as learning activities, and hence, a qualitative analysis is of immense importance to capture a concept holistically.

5.1 The concept of Extramural English (EE)

Whereas Benson's (2011) concept of language learning beyond the classroom as well as the models presented previously in this thesis encompass both language contact in and outside of a classroom setting, the present study zooms in on language learning outside an educational context exclusively. For this purpose, Sundqvist's (2009, p. 24) concept of Extramural English (EE), meaning "English outside the walls" (Sundqvist, 2009, p. 24; Sundqvist & Sylvén, 2016, p. 6), has been adopted. To be more precise, the term Extramural English refers "to the English that learners come in contact with or are involved in outside the walls of the classroom" (Sundqvist & Sylvén, 2016, p. 6). Hence, Sundqvist's (2009) definition heavily draws on the locational dimension to describe the concept of EE. Later, Sundqvist and Sylvén (2016, p. 6) elaborate on this definition by stating that apart from the fact that EE takes place outside of school, the English contact needs to be voluntarily induced by the learners or other private contacts and must not be initiated by an educational instance. Nevertheless, it is not deniable that students might participate in activities out of self-imposed pressure. Similarly, students might engage with the English

language in their spare time without the active intention to learn but rather to enjoy the activity (e.g., watching TV in English not to learn the language but out of content-driven interest), and at the same time they might consciously use their private English exposure to improve their language skills (Sundqvist & Sylvén, 2016, p. 7). For instance, EE can include incidental acquisition of new vocabulary items when watching TV out of interest, or it encompasses the active effort taken by a student when reading an English book and recording unknown vocabulary items. Hence, EE comprises both intentional and incidental language learning (Sundqvist & Sylvén, 2016, p. 12). Moreover, it is important to note that EE “does not necessarily require any action (in its literal sense) on part of the learners; the doing of an activity could simply mean that a learner is exposed to English” (Sundqvist & Sylvén, 2016, p. 11). Therefore, an EE activity does not require active or conscious involvement. To sum up Sundqvist and Sylvén’s (2016) points, EE can be characterised as follows:

EE is:

- (a) located outside of an educational context
- (b) not teacher-initiated
- (c) encompassing both intentional and incidental English contact and learning
- (d) both passive and active

For the purpose of contextualising their concept, Sundqvist and Sylvén (2016, p. 10) propose a new model of the broader field of L2 English learning (see Figure 5.5) with alphabetically ordered clouds (A-H) standing for different exemplary learning scenarios, such as students learning a language in class because it is obligatory in the curriculum (G) or a person learning a language by using a learning app because he/she is interested in the language (C).

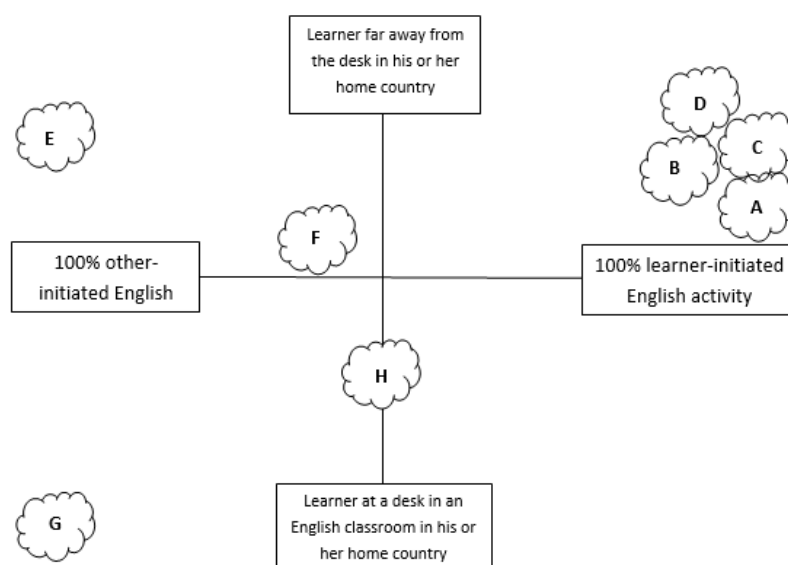


Figure 5.5: Model of L2 English learning (Sundqvist & Sylvén, 2016, p. 10)

While the vertical axis represents the dimension of location, with one end being at a desk in a classroom and the other pole being as far away as possible from this position, the horizontal axis depicts “the learner’s driving force for learning English” (Sundqvist & Sylvén, 2016, p. 11). The right end of the horizontal axis is assigned to English activities of the learner, which have resulted from his/her autonomous decisions only. On the left end of the horizontal axis, activities are 100% other-initiated, meaning that the learner does not initiate the learning experience at all. An exemplary scenario of an other-initiated English activity is if a teacher hands out a grammar task in class and asks the students to fill it in individually and silently. However, Sundqvist and Sylvén (2016, p. 11) highlight that an other-initiated activity does not exclusively refer to teacher-initiated activities but can instead also be triggered by parents, tutors, or friends. Although this appears to be a logical inference, it could prove difficult to differentiate voluntarily chosen activities from activities induced by the private surroundings of the language learner, since the learner might not even be consciously aware of or not willing to tell who the driving force behind an action was. Moreover, private external and learner initiation might be synergetic or overlapping.

Consequently, while Sundqvist and Sylvén’s (2016) model succeeds in illustrating the placement of EE within the dimensions of physical location and learner autonomy (as visible in Figure 5.6), more dimensions describing language learning beyond the classroom might need to be considered to avoid losing descriptive richness.

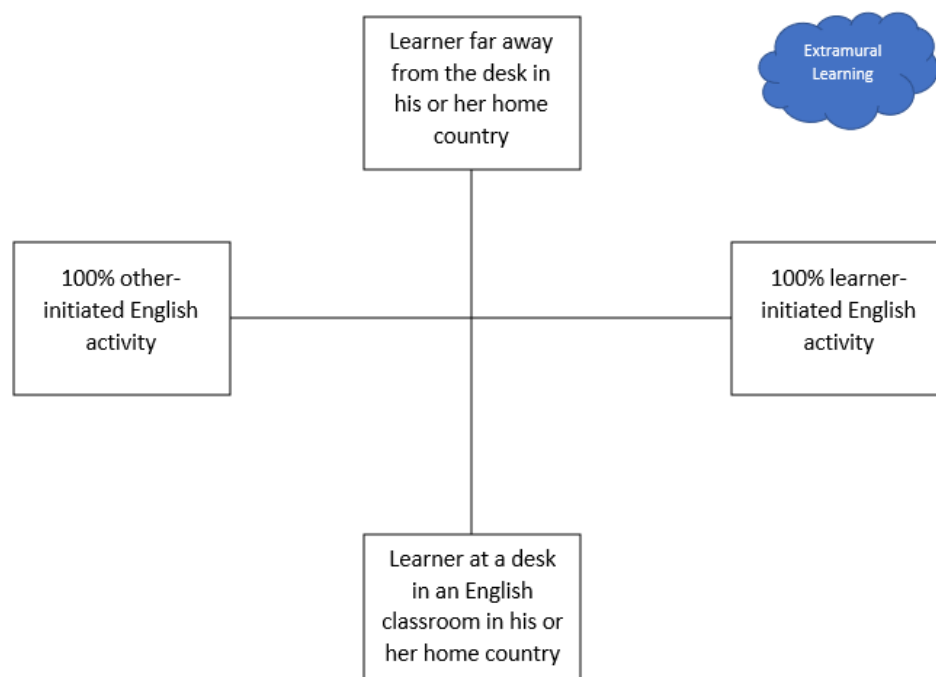


Figure 5.6: Classification of Extramural learning situations on basis of Sundqvist & Sylvén (2016)

Hence, for the purpose of contrasting and comparing concrete learning situations in more detail, the adapted model introduced in Figure 5.7 might be more effective. This figure

illustrates how the concept of Extramural English can be depicted within the area of language learning beyond the classroom.

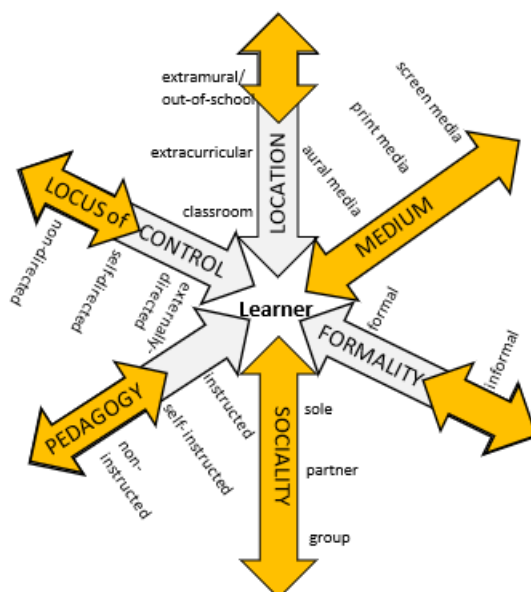


Figure 5.7: Model of Extramural English

As visualized through the yellow arrows, one can see that EE takes place outside of an educational context, uses various medias and is not directly related to assessment. However, as argued above by Sundqvist and Sylvén (2016, p. 6), self-imposed pressure might influence students EE behaviour, which is why the dimension of *formality* is not solely reduced to informal settings but allows for slight variance. EE activities can take place in various social forms, ranging from being alone to group activities. Moreover, EE activities should not be externally initiated by an authoritative figure but rather self-instructed or non-instructed. While for Sundqvist and Sylvén (2016, p. 6) EE learning activities may be initiated by private contacts other than the learner, the study at hand has limited the scope of EE to either self-instructed or non-instructed activities. Similarly, *locus of control* can be either non-directed or self-directed.

5.2 Comparison of findings

Following the clarification of the conceptualization of EE for this doctoral study, this section attempts to give an overview of the research field of EE by reviewing 115 empirical studies, which can be found in Table 5.1 below. While not all studies use the concept of EE (Sundqvist, 2009) as described above but base their research on related concepts, all of them are subsumed under the label EE in the following.

1996	- Littlewood & Liu - Pickard	1997		1998	- Yap
1999	- Freeman - Suh et al.	2000		2001	- Pill
2002	- Chan, Spratt & Humphreys - Lamb - Spratt, Humphreys, Chan - Bonnet	2003		2004	- Hyland - Lamb (a) - Lamb (b) - Pearson
2005	- Shen et al. - Yi	2006	- Sylvén - Miller & Hegelheimer	2007	- Berns, De Bot & Hasebrink - Knight
2008	- Ranalli	2009	- Sundqvist - Chusanachoti - Grau - Mukundan, Khojasteh & Pearson - Arnold	2010	- Inozu, Sahinkarakas & Yumru - Kuppens - Lefever - de Haan, Reed & Kuwada - Greenberg et al.
2011	- Veespoor, De Bot & Van Rein - Aniol - Kalaja et al. - Kuure - Lai & Gu - Lehtonen - Moncrief - Broemer & Virkkunen-Fullenwider - Olsson	2012	- Persson & Prins - Sylvén & Sundqvist - Muñoz - Sockett & Toffoli - Reinders & Wattana - Manfred	2013	- Nilsson - Barbee - Bunting & Lindström - Ekşi & Aydin - Lindgren & Muñoz - Sockett - Sundqvist (a) - Sundqvist (b) - Ferdous
2014	- Bengtsson - Jakonen - Liu - Sockett - Sundqvist & Sylvén	2015	- Sundqvist & Wikström - Batluk - Kaur - Lai, Zuh & Gong - Olsson & Sylvén - Trinder - Koivistoinen	2016	- Brevik - Cabot - Cole & Vanderplank - Lai - Lai, Yeung & Hu - Nightingale - Olsson - Liu, Lin & Wiley - Fallkvist
2017	- De Wilde & Eyckmans - Hannibal Jensen - Jóhannsdóttir	2018	- Peters - Lai, Hu & Lyu - Lai & Zheng	2019	- Sylvén - Avello, Camacho-Miñano, Urquia-Grande & del Campo

	- Elisson - Kusyk - Tran - Trinder - Hahn - Arnbjörnsdóttir		- Lancaster - Lee & Dressman - Mirmán Flores & García Jiménez - Orhon - Hlebnikovs - Ingvarsdóttir & Jóhannsdóttir - Sylvén		- Peters et al. - Puimège & Peters - Martínez Agudo - Sundqvist - Söderqvist - Lee a) - Lee b) - Lyrigkou
2020	- Rasztovits - Schwarz - Busby	2021	- Sylvén - Bergström, Norberg & Nordlund - Nihayra et al. - González-Fernández	2022	- Schurz, Coumel & Hüttner - Lee & Taylor - Sylvén & Löwenadler

Table 5.1: Empirical studies on EE sorted by publication year

Without making a claim for completeness, Figure 5.8 depicts the development of these published empirical EE studies between the years of 1996 and 2022. As can be seen, the research field experiences an overall growth with two peaks of 10 empirical studies in 2018 and 2019. Hence, this PhD thesis contributes to a highly recent and vivid research field.

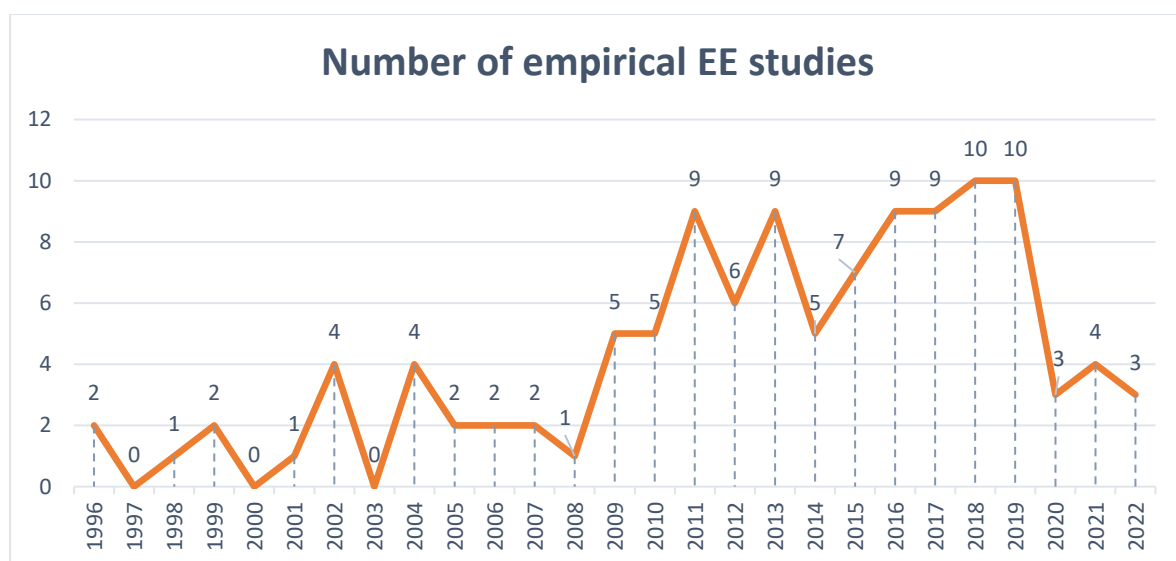


Figure 5.8: Number of empirical studies on EE from 1996 to 2022

Looking at the research field regarding the distribution of research designs reveals that there is a slight trend towards quantitative studies on EE behaviour, with 45 solely quantitative research studies in the pool of papers reviewed (see Figure 5.9).

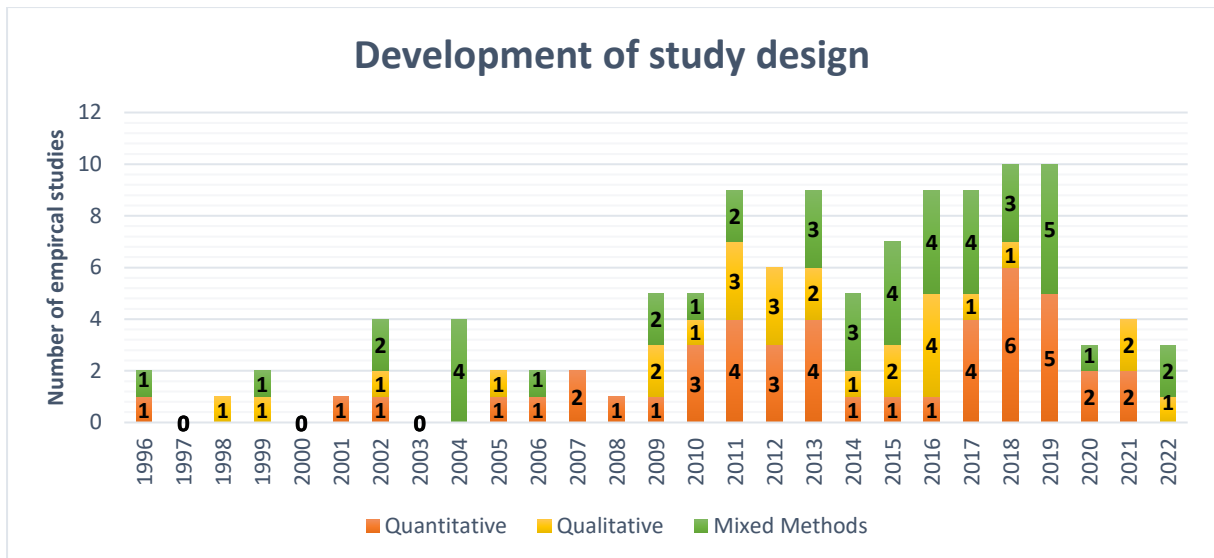


Figure 5.9: Development of study design from 1996-2022

However, quantitative studies are closely followed by 43 mixed methods studies and 27 qualitative research projects, implying a comparably balanced distribution of research designs in the research field. Nevertheless, a critical reflection of the distribution of research designs in the last years (2017-2022) indicates that there still is a need for more purely qualitative studies in the field of EE, hopefully adding depth of analysis and a clearer picture of what is done in extramural settings. This picture is also found the context of Austria, where only quantitative or mixed methods studies have been conducted so far (e.g., Schwarz, 2020; Trinder, 2015; Trinder, 2017).

A potential reason for the smaller number qualitative studies could be a lack of new methodological approaches for the investigation of EE practices. Since extramural English activities take place in participants' spare time and are characterized by their private nature, it remains difficult to investigate participants' practices in depth without being intrusive. Zooming in on the data collection methods in the studies reviewed, most researchers have relied on self-reported data in the form of questionnaires or surveys (86 studies), as can be seen in Figure 5.10.

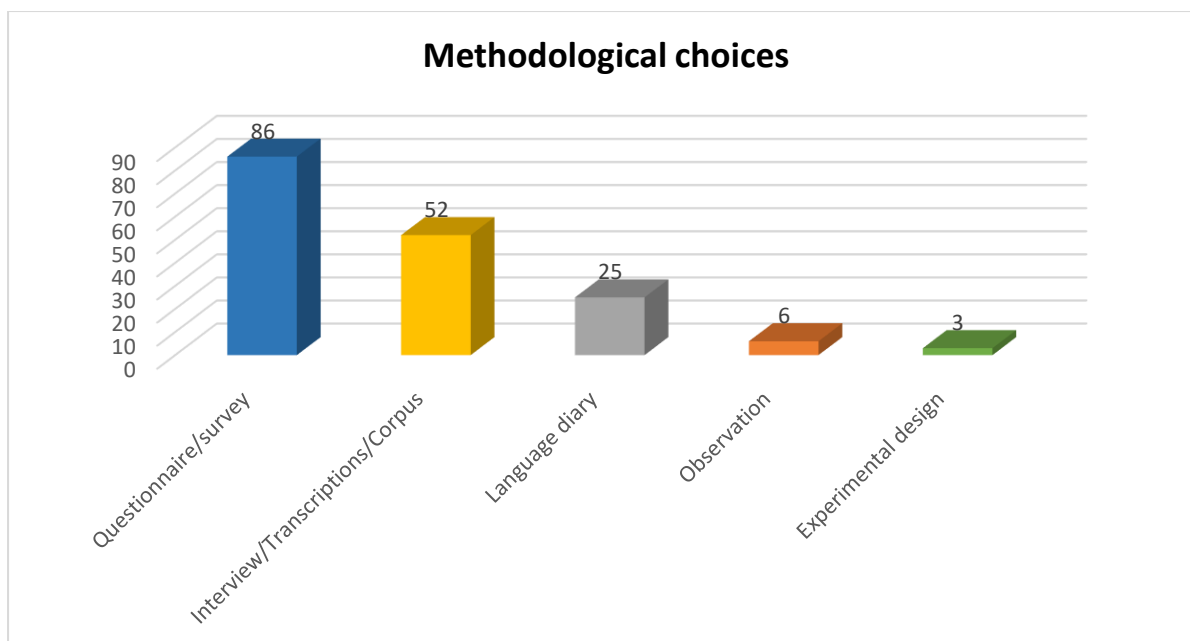


Figure 5.10: Methodological choices of empirical EE studies

While some researchers base their findings solely on questionnaire data, other studies use a combination of several tools. Building on Freeman’s (1999) work, Sylvén (2006) initiated the use of language diaries as an independent or complementary research tool. However, up to now, language diaries have been mainly used in Northern European countries, such as Sweden (Olsson, 2011; Olsson, 2016; Sundqvist, 2009; Sundqvist & Wikström, 2015; Sylvén, 2019a) or Finland (Broemer & Virkkunen-Fullenwider, 2011; Jakonen, 2014; Koivistoinen, 2015). Together with research studies from other countries using language diaries (Chusanachoti, 2009; Kuure, 2011; Schwarz, 2020), 25 research studies have relied on language diaries for their data collection, making it the third most popular research instrument. The second most popular data collection tool to investigate EE is the interview. Overall, 52 studies have conducted various types of interviews (e.g., Bunting & Lindström, 2013; Cabot, 2016; Grau, 2009; Hyland, 2004; Kaur, 2015; Lee & Dressman, 2018; Lefever, 2010; Muñoz, 2012; Pickard, 1996; Schurz et al., 2022; Schwarz, 2020).

In addition to questionnaires, interviews, and language diaries, i.e., research tools that mainly rely on self-reported data, observations and experimental designs have been used as well. However, as visible in Figure 5.10, observations and experimental designs are much less frequently used in the sample of 115 studies analysed. However, one has to consider that only experimental studies with a clear link to vocabulary knowledge have been included in the literature reviewed. Consequently, a higher number of experimental studies on EE might exist with a different focus. A more detailed graphic of the different methods used in the studies reviewed can be found in appendix 5.1.

Another interesting aspect concerning the research field of EE is a clear preference of research conducted at the level of tertiary (54 studies) and secondary education (47 studies) as can be seen in Figure 5.11.

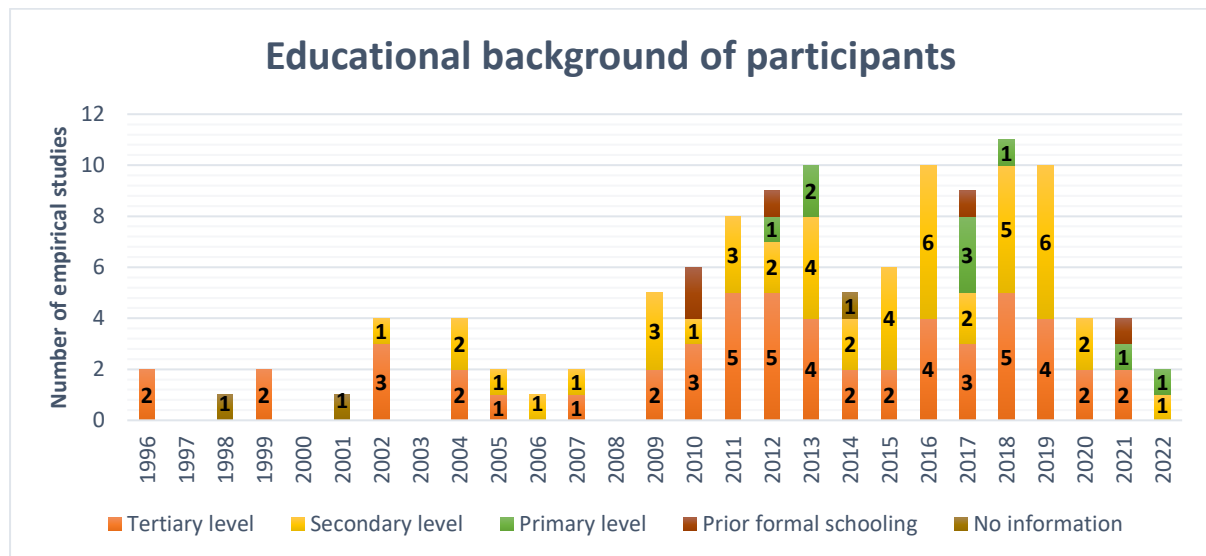


Figure 5.11: Educational background of participants in empirical EE studies

On the level of primary education, only 9 studies were conducted, and before the entrance into formal education even fewer, namely 5 studies, can be found. Hence, this is an apparent research gap to be addressed by future studies. Similarly, in the context of Austria, 7 studies on EE practices have been published, all being based in a secondary and/or tertiary educational setting (Hahn, 2017; Pickard, 1996; Rasztoovits, 2020; Schurz et al., 2022; Schwarz, 2020; Trinder, 2015; Trinder, 2017).

Regarding EE research in other countries, it can be observed that Sweden with 25 studies on EE in various contextual settings is clearly the world's best researched country in terms of EE, followed by China (14 studies), Spain and Finland (8 studies each), and Belgium (6 studies). Austria is ranked on place 5 with 7 empirical EE studies. Figure 5.12 below shows the countries already examined in terms of EE and the extent of that research.

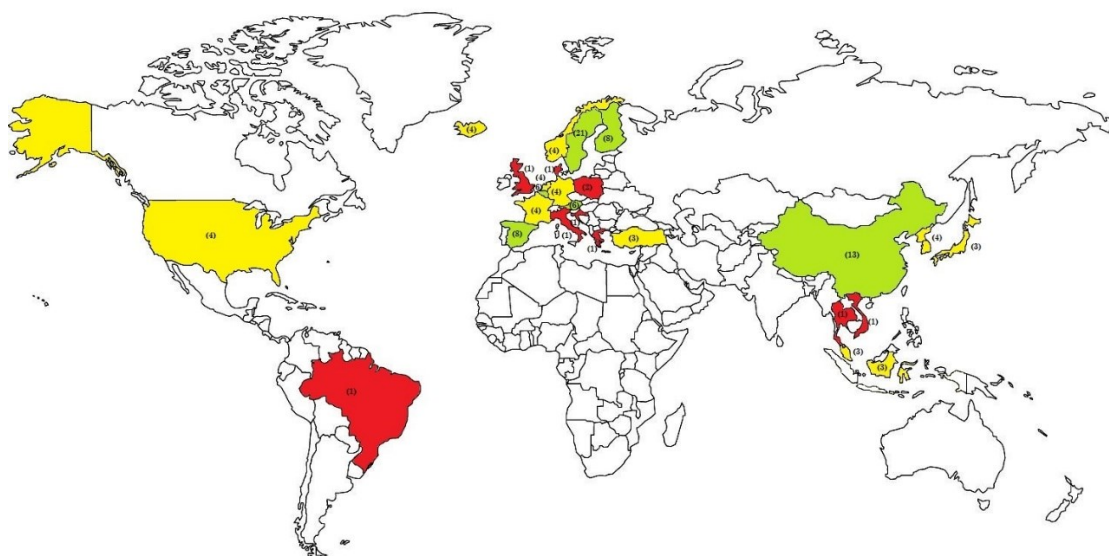


Figure 5.12: Extramural English Worldmap

Countries highlighted in green have been investigated the most (6-25 studies), followed by countries in yellow (3-5 studies). Areas marked in red have initiated research in the area of EE, with a maximum of 2 empirical studies, while white regions have not been researched at all. For more detailed information see the larger map with inserted numbers in appendix 5.2.

After this general overview of the field of Extramural English, the next section will pay more attention to similarities, differences, and trends to be found in the reported findings of the 115 studies considered. Studies conducted preceding formal English education will be reviewed first, before moving on to studies on primary school children, secondary school children and finally studies conducted at tertiary level, since a comparison across educational levels appeared to be less expedient.

5.2.1 EE preceding English formal instruction

The first study on EE activities of children aged between 7 and 8 was conducted by Lefever in 2010 in Iceland. A total of 182 children in 5 different schools were tested on their abilities in reading, listening and communication before receiving any formal English instruction (Lefever, 2010, p. 3). All testing methods used were based on the Cambridge examination for young learners, intended for children who had received between 100 and 175 hours of formal English instruction (Lefever, 2010, p. 3-4). Results show that children had a fairly good command of their listening skills, with on average 68% correct answers on the listening test (Lefever, 2010, p. 6). Regarding reading, it was found that students were in the process of “attaining basic literacy”, with an average of 50% correct answers (Lefever, 2010, p. 6). However, one could see a decline of correct answers on the reading test the more difficult the exercises were (Lefever, 2010, p. 6). To test communication skills, 51 children were interviewed by two researchers completing several tasks, such as retelling a picture story or

engaging in basic informal conversations on hobbies or pets (Lefever, 2010). While 24% of the children were able to respond fluently in English, 27% were able to have a simple conversation with prompting from the researcher. Almost half of the learners (49%) were at least able to understand simple English but could not use it actively to communicate (Lefever, 2010, p. 8). Hence, it can be concluded that a large number of children were able to participate in simple English conversations before formal English education (Lefever, 2010, p. 8). While no gender differences were found for reading and listening, a clear gender difference was detected for communication skills, with boys outperforming girls (Lefever, 2010, p. 8).

To locate potential reasons for childrens' successful use of the English language preceding formal instruction, Lefever (2010) conducted interviews with the parents of the ten most successful children, revealing that all children were heavily engaged in various extramural activities. These activities ranged from communication in English with friends (reported by 6 parents) to watching English TV shows or films (reported by all parents), listening to English songs on various media (reported by all parents) and playing computer games in English (reported by 8 parents). While online contact with English was frequent, print media, such as books or magazines were scarcely used by these children (Lefever, 2010, p. 14). Unfortunately, however, EE contact of less successful children was not investigated, eluding direct comparison.

Nevertheless, similar findings are reported by De Wilde and Eyckmans (2017, p. 677) when investigating 30 Flemish children preceding formal English education. Using comparable tools as Lefever (2010), namely the Cambridge examination for young learners - Flyers test and the Peabody Picture vocabulary test (Dunn & Dunn, 2007), participants' abilities in the skills of reading, listening, writing, speaking and receptive vocabulary were tested. Results revealed that there was a large range of test results on all 4 tests. Hence, participants performed very differently on the tests, with a bimodal distribution indicating a large gap between students in the skills of listening and speaking (De Wilde & Eyckmans, 2017, p. 680). However, in general, the mean scores for the receptive skills were higher than for productive skills, with a mean of 16.13 out of 25 for listening and 66.20 out of 108 for receptive vocabulary (De Wilde & Eyckmans, 2017, p. 679). Since reading and writing, i.e., a receptive and a productive skill, were tested together, these skills remain difficult to be placed. Nevertheless, one can infer from De Wilde and Eyckmans' (2017, p. 680) observation that "many children [have] a low score and only few children [obtain] a high score" in the combined writing and reading test, and so it behaves similar to the productive aspect of speaking. Therefore, these researchers counted this measure as productive measure.

Despite the relatively small sample size (n=30), De Wilde and Eyckman (2017) opted for a more quantitative approach than Lefever (2010) to measure students' extramural exposure to English by administering a questionnaire to both parents and students. Unfortunately,

the parental information on EE activities was excluded in the published paper. As such, the reports given by students could not be compared to the parents' judgements, which would have allowed some countermovement against the self-report character of the data collected. Moving on to the findings of this study, De Wilde and Eyckman (2017) found that digital EE activities were common, which is in line with other studies on children preceding formal education (Kuppens, 2010; Lefever, 2010; Persson & Prins, 2012). For example, children preferred to watch English TV with L1 subtitles, play English computer games or to use English for other computer activities. Another very popular activity across all studies was listening to English music (De Wilde & Eyckmans, 2017; Kuppens, 2010; Lefever, 2010; Persson & Prins, 2012). In her study of 374 Flemish children preceding formal English instruction Kuppens (2010, p. 73) even reports that listening to music was by far the most popular EE activity, as 90% of the participants said that they would be listening to English music at least 3 times a week. Similar to Lefever's (2010) findings, students of De Wilde and Eyckman's (2017) study reported to use English print media, such as books or magazines, rather seldom, but participated in spoken interaction. Despite obvious similarities between De Wilde and Eyckmans' (2017) findings and other studies (Lefever, 2010; Persson & Prins, 2012), the results on EE exposure need to be interpreted with caution, since there was no option to indicate whether an activity was not done at all. The lowest category available was 0-30 minutes participation in an EE activity per day, resulting in a potential exaggeration of EE activities and the amount of time spent on them.

Moreover, the quantitative approach used by De Wilde and Eyckman (2017) can be regarded as controversial considering the sample size ($n=30$). Nonetheless, it allows for a statistical investigation of relationships between the variables measured. A Kruskal-Wallis test revealed that the amount of time spent on gaming and other computer activities had a positive impact on all test scores but listening (De Wilde & Eyckmans, 2017, p. 682). Similar results are reported by Kuppens (2010, p. 74) with participants involved in gaming performing significantly better on two translation tests. Interestingly, in Kuppens's (2010) study, boys participated more frequently in playing computer games [$\chi^2(4) = 84.638$; $p < 0.0001$] as well as other computer activities [$\chi^2(4) = 13.106$; $p = 0.01$] than girls, implying that boys might benefit more from these EE activities (Kuppens, 2010, p. 74). In contrast, Kuppens (2010, p. 77-78) reports that girls watched significantly more TV or movies with subtitles than boys ($p < 0.01$), an activity which also had a significantly positive effect on scores on both translation tests used in her study. Hence, girls and boys at this stage appear to have different consumer behaviours from which they profit in different ways. Another positive relationship was found between the time spent on listening to English music and the Dutch-English translation test scores ($p < 0.05$) (Kuppens, 2010). Similarly, De Wilde and Eyckmans (2017, p. 682) found no significant relationship for listening to music or watching TV, which might be however partly attributed to the small sample size analysed in their study.

Persson and Prins (2012) partially confirm the findings of the studies presented above for the context of the Netherlands. However, their study compares the development of 178 children receiving early English instruction to 33 children receiving no instruction, making their results only partially applicable to the context of EE before formal education (Persson & Prins, 2012, p. 178). Moreover, while again watching TV and playing English computer games are the most prominent EE activities, no difference is made between the EE activities of the children receiving early instruction and the control group receiving no formal instruction although type and amount of EE exposure vary considerably (Persson & Prins, 2012, p. 6).

Looking at their findings, Persson and Prins (2012, p. 7) strongly emphasize that the group with formal English instruction significantly outperformed the group with no instruction on both receptive vocabulary [$F(1,145) = 4.40, p < .05, r = .17$] and grammatical knowledge [$F(1,145) = 5.04, p < .05, r = .18$] tested. Yet, it needs to be mentioned that also the group receiving no formal English input made a small but existent progress over the one-year-period investigated. Regarding receptive vocabulary knowledge, measured again by the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007), a gain of 3 points (14.2-17.2 points) was visible, while for receptive grammatical knowledge only a smaller growth of 1.2 points (5.2-6.4 points) was visible (Persson & Prins, 2012, p. 7). Opposed to these numbers, students receiving English instruction made a progress of 9 points (12.5-21.5 points) for receptive vocabulary knowledge and 3.1 points (6.8-9.9 points) for receptive grammatical knowledge (Persson & Prins, 2012, p. 7). Unfortunately, there is no information whether the gain of the control group was also significant and whether EE activities had a potential impact. Nevertheless, participants in the early English instruction programme benefitted the most from the quality of input in class and watching English TV shows for monolingual children (Persson & Prins, 2012, p. 7).

A related study by Nihayra et al. (2021, p. 5) of 262 10-year-old Dutch children investigated the influence of EE exposure on both oral and written receptive vocabulary in two opposed groups, namely children with and without formal English instruction at school. Using an EE questionnaire, the authors revealed that there was a significant difference between the two groups regarding overall EE exposure: 22.5% of the children with formal English instruction at school and 50.9% without formal English instruction were never involved in EE activities (Nihayra et al., 2021, p. 7). Moreover, they found that especially EE activities performed within the family and EE actions related to entertaining media had a greater impact on students' receptive vocabulary when they had no formal instruction (Nihayra et al., 2021, p. 7). This shows not only that more in-depth investigation of the influence of explicit and

implicit learning on vocabulary knowledge is necessary but that, additionally, the interplay between formal and informal learning needs to be examined.⁵

5.2.2 EE at primary school level

Similar to Persson and Prins (2012), Hannibal Jensen (2017, p. 6) investigated the influence of EE activities on the receptive vocabulary knowledge of 107 Danish primary school children, 49 being early starters in regard to learning English and 58 being late starters. Conversely to the findings by Persson and Prins (2012, p.7) regarding children receiving early English instruction opposed to children receiving no formal instruction, early starters in Denmark had no advantage regarding receptive vocabulary knowledge over late starters and performed even weaker on the Peabody Picture vocabulary test (Hannibal Jensen, 2017, p. 10). This observation was also made by Jóhannsdóttir (2017, p. 57) in her study of 416 Icelandic primary school children, where “onset of English instruction was not a strong predictor for lexical proficiency”. In addition to starting age, the score on the receptive vocabulary test was also predicted by the amount of gaming children were involved in as well as by their gender (Hannibal Jensen, 2017, p. 12). This reflects Hannibal Jensen’s (2017) findings regarding EE. While girls spent most time on listening to music (115.4 min/week), followed by watching English TV (69.8 min/week) and gaming (46.8 min/week), boys spent most time on gaming (235.1 min/week), similarly followed by TV (120.9 min/week) and music only on third place (78.8 min/week). As can already be inferred by the amount of time spent on these three activities mentioned in brackets above, significant differences between learners with different gender were found for gaming time and time spent on music (Hannibal Jensen, 2017, p. 8). Furthermore, boys, who were more involved in gaming, were more successful than girls regarding the receptive vocabulary test. Consequently, girls in primary school, in contrast to female learners preceding formal education (see Kuppens, 2010), were not able to compensate for their lack of gaming through involvement in other EE activities.

A similar picture regarding extramural activities in primary school was described by Bunting and Lindström (2013) and Elisson (2017). Both interview studies, the former involving 47 Swedish and the latter 50 Swedish primary school children, report that their participants mentioned watching TV as crucial EE activity as well as gaming (Bunting & Lindström, 2013, p. 209; Elisson, 2017, p. 9). In agreement with Hannibal Jensen’s (2017) study, a gender difference regarding the importance of gaming was found (Bunting & Lindström, 2013, p. 209). While girls mentioned TV or watching movies as the most prominent EE source, boys mentioned digital games as their favourite EE activity (Bunting & Lindström, 2013, p. 209).

⁵ Since this study focuses solely on the impact of EE on different aspects of vocabulary knowledge, a main emphasis on informal language learning has been chosen. However, additional data related to vocabulary learning of the same participants in formal settings has been collected, which can not be presented within the scope of this study. However, the relation between formal and informal learning settings will be examined in more detail in further publications.

However, when being asked directly about playing computer games, 74% of the girls reported that they were involved in gaming. Yet only 11% of the girls were playing multiplayer online games, which require written and spoken interaction with other language users, while 85% of the boys preferred playing this type of games (Bunting and Lindström, 2013, p. 210). Instead, girls chose to play computer games based on written English language only (Hannibal Jensen, 2017, p. 11). Hence, the difference regarding gaming in primary school might have been narrowed regarding time spent on gaming but apparently the type of game played has a crucial impact as well. This is a fact, which has been confirmed by several studies for secondary school students already (e.g. Sylvén & Sundqvist, 2012; Sundqvist & Sylvén, 2014). However, it remains questionable whether girls are able to compensate their minor engagement in gaming by participating more intensively in other activities, as implied by Kuppens (2010) for children entering formal education. Hence more studies focusing on gender-specific EE use at primary school level need to be conducted.

While the two interview studies at the level of primary education mentioned above (Bunting & Lindström, 2013; Elisson, 2017) agree on most findings, such as the importance of watching TV or movies in English, one apparent difference can be detected. While participants of Bunting and Lindström (2013) assigned little importance to listening to music, interviewees of Elisson (2017, p. 9) mentioned listening to music as the second most frequent English activity they were involved in extramurally. Looking at other, more quantitative studies (Jóhannsdóttir, 2017; Lindgren & Muñoz, 2013) reveals that listening to English music tends to be a rather popular EE activity amongst primary school children; a behaviour that appears to be retained from EE input prior to formal education (De Wilde & Eyckmans, 2017; Kuppens, 2010; Lefever, 2010; Persson & Prins, 2012). In Lindgren and Muñoz' (2013) study of 865 primary school children from various countries, listening to music was even the most frequently mentioned EE activity, with an average of 1.75 hours spent on it per day, and was additionally exercising the most influence on success in listening and reading activities, followed by watching films or TV and again playing computer games.

Regarding external factors influencing primary school childrens' listening and reading skills, age, gender (Lindgren & Muñoz, 2013, p. 118-119) and parents' professional use of English, as well as the cognate relatedness of the L1 and English (Hannibal Jensen, 2017) were identified as central factors. Additionally, reading ability appears to be linked to parental education. Zooming in on the productive skill of speaking, Lee and Taylor (2022) investigate the impact of EE engagement and positive psychology on 160 primary school childrens' willingness to communicate in English. Using a mixed methods approach based on survey data and semi-structured interviews, their findings reveal that childrens' willingness to communicate in English in and outside class is both predicted by their EE engagement and aspects of positive psychology (Lee & Taylor, 2022, p. 10). Consequently, they propose that English language teachers should use EE integrated intervention programmes in form of a

so-called “EE challenge” (Lee & Taylor, 2022, p. 13). To complete such a challenge, the learner needs to identify his/her learning objectives (e.g. improving English reading) and needs to choose suitable resources and EE activities to reach this goal. However, choice of objective, material and activity should be left to the learner. The teacher, rather than being the main instructor, should provide the learners with affective, behavioural and cognitive support.

Another interesting aspect to be considered in relation to EE behaviour is the type of education students are involved in. By comparing 156 CLIL students to 162 non-CLIL students in the context of primary and secondary education in Spain, Lancaster (2018, p. 98) establishes another engaging research area within the field of EE. While no differences in the amount of EE exposure between CLIL and non-CLIL students were found at primary school level, significant differences were detected at secondary school level ($p < 0.006$) (Lancaster, 2018, p. 101). Tendentially, CLIL students in secondary school were involved in more EE activities, such as reading English books or magazines, internet use in English or listening to English music, than their EFL counterparts (Lancaster, 2018, p. 103). Moreover, at secondary school level, CLIL students performed significantly better on vocabulary, listening and pronunciation tests than non-CLIL students, which was related to their increased extramural exposure to the target language (Lancaster, 2018, p. 103). Yet, more studies need to be conducted at the level of primary education to determine which factors influence language learning the most.

5.2.3 EE at secondary school level

Similar to Lancaster (2018), Sylvén (2006, p. 49) reports significant differences regarding the amount of EE exposure of CLIL (13h/week) and non-CLIL students (9h/week) at secondary school level, with CLIL students spending more time on reading English books and surfing the internet in their spare time. These findings are in accordance with Olsson and Sylvén (2015), who repeatedly investigated EE practices of 184 CLIL versus 84 non CLIL students. CLIL students significantly ($p = 0.024$; $p = 0.006$) outperformed non-CLIL students regarding both amount of EE practices and time spent on them, with CLIL students spending 7.65 hours and the control group only 5.6 hours a day on EE activities (Olsson & Sylvén, 2015, p. 90). Digging deeper into the learners’ objectives for EE activities, Sylvén’s (2019a, p. 119) qualitative case study one CLIL and one regular student shows that the CLIL student rather conceptualized the L2 from an active point of view as something to be used, while the non-CLIL student foregrounded learning. Similar results are additionally reported by Sylvén (2019b) and Martínez Agudo (2019) in Spain.⁶

⁶ Nonetheless, it is important to consider that great variance between individuals was found in all of the studies reported (Sylvén, 2006; Olsson & Sylvén, 2015; Lancaster, 2018; Martínez Agudo, 2019; Sylvén, 2019b). Hence, individual EE behaviour might be very divergent from the average reported in these studies, which highlights the importance of more qualitative approaches to EE of CLIL vs. non-CLIL students in the future.

While an overall agreement exists regarding differences of EE behaviour of CLIL versus non-CLIL students, the intensity of involvement in EE activities reported in Olsson and Sylvéns' (2015) study is not reflected in other studies (Lai, Zuh & Gong, 2015; Martínez Agudo, 2019; Olsson, 2011; Sundqvist, 2009; Sundqvist & Sylvé, 2014; Sylvé & Sundqvist, 2012). Assuming that the time reported to be spent on EE activities per day in Olsson and Sylvéns' (2015) study can be multiplied by seven for the whole week, not considering that weekends might offer even more time for spare time activities of all kinds, results in 55.55 hours spent on EE activities by CLIL students and 39.2 hours by non-CLIL students per week. However, in comparison to other studies (Lai, Zuh & Gong, 2015; Martínez Agudo, 2019; Olsson, 2011; Olsson, 2016; Sundqvist & Sylvé, 2014; Sylvé & Sundqvist, 2012; Sylvé 2019), this amount of time appears to be rather extensive, also taking into account that a week has 168 hours without consideration of sleep or other daily activities and time spent at school.

Similar results were only found by Sylvé (2019a, p. 161) with 55.6-66.73 hours spent on EE activities per week by CLIL students and 39.32-29.87 hours per week for non-CLIL students. For Austria, Schwarz (2020, p. 218) partially confirms Olsson and Sylvéns' (2015) and Sylvé's (2019) findings regarding time spent on EE activities by non-CLIL students, with her students devoting on average 4 hours and 6 minutes on EE on weekdays and 5 hours and 35 minutes on weekends. This amounts to an average of 31.75 hours per week spent on EE activities. Another study in the context of Spain on the difference between 156 CLIL and 162 non-CLIL students arrives at less time spent on EE activities by both CLIL and non-CLIL students, with 22.08 hours per week for CLIL students and 17.75 hours per week for non-CLIL students (Martínez Agudo, 2019, p. 76). Other studies solely investigating non-CLIL students appear to be in accordance with Martínez Agudo's (2019) results. For example, a study on Swedish secondary school students by Sundqvist (2009, p. 116) determines that students spent a mean time of 18.4 hours and a median time of 15.9 hours per week on EE activities. Rather similar results were found by Olsson (2011, p. 38) also investigating Swedish secondary school pupils with approximately 20.3 hours invested in extramural activities per week. In contrast, other studies report lower results with 9.4 hours per week (Sylvé & Sundqvist, 2012, p. 310), approximately 8.6 hours per week (Lai, Zuh & Gong, 2015, p. 289) and 7.2 hours per week (Sundqvist & Sylvé, 2014, p. 10).

These differences of time spent on EE activities amongst CLIL students might be attributed to the different settings of the studies, with all studies set in Sweden reporting high exposure unlike Martínez Agudo's (2019) Spanish data. Yet, the observed discrepancies regarding non-CLIL students remain largely unaccounted for, since most studies examining time spent on EE activities are from Nordic countries with Sweden being the most prominent nationality. Hence, more studies like Lai, Zuh and Gong (2015), Martínez Agudo

(2019) or Schwarz (2020) are needed to enable explanations for divergent results and to allow for international comparison.⁷

So far, only two additional influencing factors on EE behaviour at secondary school level apart from the geographical context were researched. Regarding biographical conditions of participants, Berns, de Bot and Hasebrink (2007) claim that the level of English proficiency of parents and siblings is significantly related to the amount of extramural exposure on the basis of self-reported data of 2200 students in several countries. However, despite the impressive number of people included in the study, this result needs to be interpreted with caution, since the data collected appears to lack comparability due to different questionnaires administered in different settings. Moreover, several age groups and school types are subsumed and mixed, additionally reducing comparability. Looking at other studies, this claim seems difficult to hold as well. For example, while 87.6% of the 186 Spanish secondary students surveyed by Mirmán Flores and García Jiménez (2018, p. 291) believed that their parents exercised great influence on their extramural behaviour, the researchers could not detect any influence of parents' level of English. Moreover, other studies (e.g., Sundqvist, 2009) could not discover any links between biographical background variables and EE exposure.

In contrast, intrinsic variables, such as level of enjoyment reported for EE activities, have been shown to have a positive effect on the amount of EE exposure at secondary school level (Barbee, 2013, p. 27). In his questionnaire study of 151 secondary school students in Japan, Barbee (2013, p. 20) discovered that the EE activities pursued the most by the participants, namely listening to music with and without lyrics, using online media and watching movies with English subtitles, correspond to the actions reported to be the most enjoyable. Yet, participants named some different EE activities when being asked which activities they believe to be most effective for learning English (Barbee, 2013, p. 21). While music and TV were not only enjoyable but also believed to be beneficial for improving one's English, interaction with native speakers was considered to be the most effective spare time activity. However, no inferential statistics were calculated to further validate these descriptive results. In the same vein, Ingvarsdóttir and Jóhannsdóttir (2018, p. 89-90) report that Icelandic secondary school students have rather 'traditional' beliefs about activities that help them to improve their English skills. 77% of 703 Icelandic students aged between 15 and 16 claimed that they learned very much or much from grammar exercises and over 75% of them believed that translation exercises helped them. In contrast, more informal activities, such as playing computer games (36%) or participation in role plays (25%), were attributed much less learning potential. However, at least 59% stated that they learned very much or much from listening to English music, and 76% felt that they learned

⁷ Additional to studies exploring EE behaviour in another setting, cross-national studies might offer potential for accurate comparison across countries, assuming they consistently use the same research instruments in all settings.

much or very much from watching English videos. Moreover, 75% believed that reading English books or magazines helped (Ingvarsdóttir & Jóhannsdóttir, 2018, p. 90).

When having a look at students' preferred EE behaviour and potential approaches towards categorization reveals several interesting differences. While most studies report a preference of secondary school students for receptive EE practices (e.g., Hahn, 2017; Lai, Zuh & Gong, 2015; Lyrigkou, 2019; Rasztovits, 2020; Schwarz, 2020; Söderqvist, 2019; Sundqvist, 2009), such as watching TV or listening to music, a differentiation of EE activities via mode of input similar to Barbee (2013, p. 13) appears to be more suitable, since many activities can be practiced solely receptively but also partly productively. For instance, listening to music might be a completely receptive action or could include production when students sing along or decide to note down lyrics they particularly like. Hence, Barbee's (2013, p. 13) differentiation of three different modes of input adds distinctive dimensions. He distinguishes between (1) mixed modalities, such as watching TV or playing computer games, (2) verbal modes, such as listening to music or the radio and (3) written modes of input, as reading a book or the newspaper (Barbee, 2013, p. 2). In his study of 151 Japanese secondary school students, mixed modal EE activities were preferred over verbal EE activities followed by written activities (Barbee, 2013, p. 1). This is a pattern which can be observed in many other studies as well (Aniol, 2011; Grau, 2009; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén & Sundqvist, 2012).

Regarding the types of EE activities and their frequencies, clear trends are visible, with listening to English music being either the most practiced extramural activity (Barbee, 2013; Berns, De Bot & Hasebrink, 2007; Grau, 2009; Hahn, 2017; Ingvarsdóttir & Jóhannsdóttir, 2018; Lai, Zuh & Gong, 2015; Lamb, 2004a; Lamb, 2004b; Lyrigkou, 2019; Nilsson, 2013; Olsson, 2011; Peters, 2018; Schwarz, 2020; Söderqvist, 2019; Sundqvist, 2009) or among the most common EE practices (Aniol, 2011; Fallkvist, 2016; Sundqvist & Sylvén, 2014). This tends to be followed by consumption of video material, being regular TV programme, films, series or online video clips (Grau, 2009; Hahn, 2017; Ingvarsdóttir & Jóhannsdóttir, 2018; Lai, Zuh & Gong, 2015; Lamb, 2004a; Lamb, 2004b; Lyrigkou, 2019; Nilsson, 2013; Olsson, 2011; Peters, 2018; Schwarz, 2020; Söderqvist, 2019). Exceptions regarding these results can be found in Aniol (2011, p. 109), with Polish and Norwegian participants being rather uninvolved in English music but quite engaged in English video material, and in Barbee (2013, p. 19), whose participants reported a reverse picture, with much EE time devoted to music, while TV or movies in English are only found on the fourth (with L1 subtitles) or seventh rank (without subtitles). However, while overall video consumption is clearly amongst the most popular extramural activities, it remains difficult to differentiate between types of video input due to a focus on the terminal device of existent studies. While different input through various technical devices might have been common in the past, the same video material can be watched on several tools nowadays. For instance, children might watch *the Sandman* as regular TV programme on the television at home, on the tablet

in the car or on the parents' mobile phone. Hence, it is proposed that future studies should focus more on the type of video material consumed, such as a two-hour movie or a series lasting over several weeks, instead of concentrating on the terminal device.

While the third rank of most common EE activities already becomes more varied, many studies report that computer games are the third most popular choice (Grau, 2009; Sundqvist, 2009; Sylvén & Sundqvist, 2012; Nilsson, 2013; Barbee, 2013; Sundqvist & Sylvén, 2014; Peters, 2018). However, more studies report less prominent results on the use of English computer games (Aniol, 2011; Hahn, 2017; Hlebnikovs, 2018; Ingvarsdóttir & Jóhannsdóttir, 2018; Lyrigkou, 2019; Mirmán Flores & García Jiménez, 2018; Schwarz, 2020; Söderqvist, 2019). What becomes visible by comparing these empirical studies is that the national context of studies finding less emphasize on gaming during EE time is rather varied, whereas studies reporting much gaming time are predominantly from Northern countries. Hence, results might be country specific. Apart from differences in the time spent on playing computer games, a clear trend regarding type of games played has emerged. Boys tend to prefer multiplayer online games or physically oriented games, such as first-person shooter games. In contrast, female players choose more thought-oriented single player games, which are played offline (Greenberg et al. 2010; Sylvén & Sundqvist, 2012). However, while preferences regarding the type of game appear to soften with age, older players reported that they played more because of the competition in comparison to their younger counterparts. Instead, younger games rather appear to play games due to the fantasy needed to play them (Greenberg et al., 2010).

Although clear trends regarding the preference of EE activities are visible overall, rankings should be treated with caution for several reasons. Firstly, extramural behaviour is a strongly individualised phenomenon showing large variance from participant to participant (e.g., Lamb, 2004a; Sundqvist, 2009; Sylvén, 2006; Yi, 2005). Hence, qualitative studies, such as the case study of a 12-year-old Finnish boy by Koivistoinen (2015) or Sylvén's (2018, 2021) close investigations of individual children, are needed to perceive the individual side of EE. In a case study of Jimmy, an L1 Swedish speaking boy, Sylvén (2021) uses interviews to examine his views on gaming and differences between EE and formal English instruction. Additionally, audio recordings of Jimmy playing computer games have been collected in addition to several language tests (Sylvén, 2021). Another recent study (Sylvén, 2018) of two sisters aged 5 and 7 years shows that both children display strong self-confidence and indicators for self-selected bilingualism in English and their L1 Swedish, linked to their consumption of YouTube videos.

Secondly, studies are sometimes difficult to compare due to differences in methodology and reporting methods used. Moreover, article length does often not allow the authors to give as much insight as would be needed to make a qualified judgement of data and results. For

instance, rankings are mostly based on mean scores, without the necessary information about data distribution.⁸

However, apart from within and between study differences regarding time spent on computer games other influencing variables were identified, with the most dominant being gender. While boys spend considerable time on playing computer games in their leisure time, girls appear to show much less interest in this EE activity (Fallkvist, 2016; Grau, 2009; Greenberg et al., 2010; Hlebnikovs, 2018; Olsson, 2011; Olsson & Sylvén, 2015; Peters, 2018; Rasztovits, 2020; Schwarz, 2020; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén, 2006; Sylvén & Sundqvist, 2012; Tran, 2017). Nevertheless, girls were found to perform other EE activities much more frequently than boys, such as music (Schwarz, 2020; Sylvén, 2006), reading (Hlebnikovs, 2018; Olsson, 2011; Schwarz, 2020), watching TV (Grau, 2009), using the internet in general (Sylvén, 2006) and, more specifically, for social media activities (Schwarz, 2020; Sundqvist & Sylvén, 2014). Moreover, while tendentially little time is devoted to speaking in the spare time (Lamb, 2004a; Lamb, 2004b; Sundqvist, 2009), girls appear to spend more EE time speaking the English language (Olsson, 2011, p. 34) than boys. However, these results are not as consistent as the gender gap regarding time spent on gaming. While Sylvén's (2006, p. 50) and Schwarz's (2020, p. 226) data sets imply that girls listen more often to music than boys, Lamb (2004a & 2004b) did not find such a tendency in 219 respondents. Additionally, Grau (2009, p. 166) states that, out of her 900 German participants, girls did not spend more time than boys on listening to music in general but that girls engaged with the lyrics of the songs significantly more often. However, a study by Batluk (2015, p. 17- 19) including both sexes shows that all participants were not only consuming music passively but were working with the lyrics, singing along or listening to songs repeatedly, which fosters vocabulary uptake.

Similar disagreement can be found regarding the gender-based use of English-based internet sources. Investigating overall internet usage, Sundqvist (2009, p. 120) and Olsson (2011, p. 35) found a statistical difference to the advantage of boys. Yet, it seems as if great potential would lie in narrowing the exact subject of investigation from overall internet activities to more specific tasks performed online. For instance, Sundqvist and Sylvén (2014, p. 11) and Schwarz (2020, p. 226) both attribute more online time to girls when solely looking at social media practices done online.

Regarding watching TV, Sundqvist and Sylvén (2014) contradict Grau's (2009, p. 166) finding that girls consume more video material. The only gender difference apart from playing

⁸ As outlined in Field (2018, p. 228) and Dörnyei (2007), a mean score should only be used for interpretation of data if the data is normally distributed. If the data is not normally distributed, the median score should be presented in addition to the mean, since the median score might be the better representation of the data average in this case. Extreme scores or outliers, which are likely to occur in non-normally distributed data, might distort the mean score, while the median is not vulnerable regarding these phenomena. Consequently, rankings based on mean scores of non-normally distributed data might be misleading.

computer games agreed on by all studies included is the prevalence of female readers, which is an interesting result, especially considering that studies consistently report that few EE time is spent on reading books or magazines (Aniol, 2011; Barbee, 2013; Fallkvist, 2016; Nilsson, 2013; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén & Sundqvist, 2012; Yi, 2005). However, this finding, too, needs to be relativized, since several studies indicate that literacy behaviour amongst secondary school students has rather changed than declined (Schwarz, 2020; Sundqvist, 2009; Yi, 2005). As foreshadowed by Schwarz (2020, p. 214-219), students transferred their reading time to online sources, such as social media, instead of reading print materials, such as novels or journal articles.⁹

While the number of EE activities performed by girls mentioned above might imply that girls simply compensate for their lack of time invested in gaming with more investment in other EE activities, some studies suggests that boys also spend more time on EE activities in absolute terms (Fallkvist, 2016; Olsson, 2016; Olsson & Sylvén, 2015; Sundqvist & Sylvén, 2014). However, a similarly large number of studies report that no or no statistically significant gender difference regarding EE time can be found (Grau, 2009; Olsson, 2011; Sundqvist, 2009; Sylvén & Sundqvist, 2012). Yet, boys appear to participate in more active EE activities than girls (Fallkvist, 2016; Sundqvist, 2009).

Apart from investigating the extent and type of EE activities as outlined above, some studies examine which potential benefits increased extramural exposure might have. So far, EE has shown to have a positive influence on both intrinsic emotional factors and on language skills (Lai, Zuh & Gong, 2015; Olsson, 2011; Schwarz, 2020; Sylvén, 2006). For instance, a mixed methods study amongst 82 Chinese secondary school students by Lai, Zuh and Gong (2015) shows that the more participants engaged in EE activities, the more they enjoyed the language itself. Similar findings are reported by 152 Spanish participants in Nightingale's (2016) study on 536 secondary school students. Similarly, in Verspoor, De Bot and Van Reins' (2011) study, a positive relationship between EE exposure and positive attitude towards the English language could be detected. Moreover, amount (Sylvén, 2006) and diversity (Lai, Zuh & Gong, 2015) of extramural exposure were significantly related to participants' self-confidence.

Concerning language skills, the amount of EE shows to be positively linked to level of oral proficiency (Sundqvist, 2009), written proficiency (Olsson, 2011; Verspoor, De Bot & Van Rein, 2011) and vocabulary knowledge (Fallkvist, 2016; Olsson, 2011; Olsson & Sylvén, 2015; Peters, 2018; Schwarz, 2020; Sundqvist, 2009; Verspoor, De Bot & Van Rein, 2011) and consequently on grades (Lai, Zuh & Gong, 2015). In her PhD study involving 80 Swedish secondary school students aged between 15 and 16, Sundqvist (2009) investigated the

⁹ Another trend reported by Nilsson (2013, p. 39-40) related to reading at secondary school level is the preference of American resources over British English media.

influence of amount of extramural exposure on oral proficiency as well as receptive and productive vocabulary knowledge. By utilization of an EE questionnaire, 5 interactional speaking tests and three vocabulary tests, being the cloze test of the Swedish National Test and adapted versions of the Vocabulary Levels Test (Nation, 2001) and the Productive Vocabulary Levels Test (Nation, 2001), she was able to detect significant positive relationships between EE and both oral proficiency ($r_s=.307$) and degree of receptive and productive vocabulary knowledge ($r_s=.352$; $r_s=.354$) for male participants. However, no significant link between amount of EE and proficiency in these two skills could be found for girls (Sundqvist, 2009, p. 142), although no significant gender differences regarding oral proficiency and productive vocabulary size could be identified. Therefore, Sundqvist (2009) argues that the impact of EE might be greater for boys than for girls. Other studies however have found statistically significant relationships for both genders (Fallkvist, 2016; Olsson, 2011; Olsson, 2016; Schwarz, 2020). For instance, Olsson (2011) found that Swedish pupils who engaged more in EE were using more technical vocabulary and had a better type/token ratio than students who were less exposed. Likewise, Peters (2018) and Schwarz (2020) found statistically significant relationships between EE exposure with vocabulary knowledge for both genders. However, while Peters (2018) found positive correlations for both productive and receptive vocabulary knowledge using the VocabLab Test (Peters, Velghe & Van Rompaey, 2019), Schwarz (2020) identified only a relationship to pupils' receptive vocabulary size, measured by an adapted Yes/No test. Nevertheless, she argues that the lack of a correlation could be due to the testing method and that traces for productive growth can be found in other types of her data (Schwarz, 2020, p. 326).

Investigating these studies in more detail reveals that specific types of extramural exposure appear to be particularly fruitful for vocabulary development. While in Sundqvist (2009, p. 156) gaming is the most influential variable followed by using the internet, watching TV and reading newspapers, Peters (2018, p. 157) only found influences from watching TV ($r^2=.14$) or films without subtitles ($r^2=.10$) and using the internet ($r^2=.17$) but not for gaming. However, gaming has been shown to have a positive influence on listening, reading and vocabulary skills in several other quantitative and qualitative studies (Brevik, 2016; Cabot, 2016; Greenberg et al., 2010; Sundqvist, 2013a; Sundqvist, 2013b; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Interestingly, the amount of time spent on gaming and the amount of time devoted to reading print material appear to be directly negatively related (Sylvén & Sundqvist, 2012, p. 312). Nonetheless, the amount of time spent on reading books or magazines still appears to have a significant positive impact on vocabulary scores as well (Peters, 2018, p. 157). Considering the low amount of time devoted to reading print materials, this result is even more surprising, pointing to the importance of engagement with this type of media.

Overall, amount and type of extramural exposure appears to have a crucial influence on both receptive and productive language skills. Therefore, as qualitative studies, such as

Cabot (2016) show, out-of-school contact with the English language additionally enables students to become active agents of their language learning.

5.2.4 EE at tertiary level

In contrast to studies on extramural exposure at secondary school level, few findings regarding the exact quantity of EE are available at tertiary level. While Arnbjörnsdóttir (2017, p. 43) reports 0-3 hours per day spent on EE activities requiring listening and 0-4 hours a day on speaking activities, Littlewood and Liu (1996) and Lee (2019a; 2019b) have found little exposure to EE, with 49.4% of 77 Korean university students (Lee, 2019a, p. 772) and 62% of 71 Korean university students (Lee, 2019b, p. 119) spending less than an hour per day on extramural practices. On a more positive note, Inozu, Sahinkarakas and Yumru (2010, p. 16) report that 79% of their 309 Turkish first-year university students at least look for opportunities to use English outside of the classroom context. Nevertheless, similar to the results in Asian countries, opportunities for EE appear to be rather scarce, since still 44% report using English rarely outside of school (Inozu, Sahinkarakas & Yumru, 2010, p. 16). These findings match Cole and Vanderplank's (2016, p. 39) study investigating the amount of time spent on EE by classroom-trained English learners. However, while only 25% of the classroom-trained learners spent more than 2 hours a day on EE activities, 54% of the self-instructed learners involved in the study invest more than 2 hours per day on EE (Cole & Vanderplank, 2016, p. 39). Looking only at the extramural English experiences via digital devices of 5 tertiary level students, Sockett and Toffoli (2012, p. 142) found that most participants spent about 37 hours in 8 weeks in contact with the English language on online devices. Likewise, Lai and Zheng (2018, p. 307) reveal that 256 Chinese university students spent about 1-3 hours a week on EE solely on their mobile phones. While these studies allow a first impression concerning the amount of time spent on EE, there is a considerable backlog in comparison to the degree of exploration of frequency of EE at secondary school level. Additionally, a research gap regarding gender differences at the level of tertiary education can be located compared to research conducted at secondary school level.

Regarding the type of EE exposure, similarities and differences to prior educational levels can be detected. In comparison to studies at primary and secondary school level (e.g. Aniol, 2011; Barbee, 2013; Fallkvist, 2016; Nilsson, 2013; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén & Sundqvist, 2012; Yi, 2005), several studies at tertiary level have identified a stronger focus on reading (e.g., Chan, Spratt & Humphreys, 2002; Ferdous, 2013; González-Fernández, 2021; Hyland, 2004; Inozu, Sahinkarakas & Yumru, 2010; Knight, 2007; Lamb, 2002; Littlewood & Liu, 1996; Liu, 2014; Liu, Lin & Wiley, 2016; Mukundan, Khojasteh & Pearson, 2009; Moncrief, 2011; Pearson, 2004; Pickard, 1996; Spratt, Humphreys & Chan, 2002). For instance, Pearson (2004, p. 4) discovered that 73% of her 106 Chinese university students read English books, magazines, and newspapers, and González-Fernández (2021) listed reading as the third most popular activity, closely after watching TV and listening to

music. However, studies do not fully agree on the popularity of different print media amongst tertiary students (e.g., Ferdous, 2013; Hyland, 2004; Inozu, Shinakarakas & Yumru, 2010; Liu, 2014; Liu, Lin & Wiley, 2016; Moncrief, 2011; Mukundan, Khojasteh & Pearson, 2009; Orhon, 2018; Pickard, 1996). While Inozu, Shinakarakas and Yumru (2010, p. 16) as well as Moncrief (2011, p. 112) state that books and English novels are the most popular print medium followed by magazines and newspapers, Mukundan, Khojasteh and Pearson (2009, p. 46), Liu (2014, p. 77) and Orhon (2018, p. 7) found the reverse picture. However, overall all studies agree that English novels are rather popular amongst tertiary-level students. Moreover, with academic texts, a new medium has gained popularity at tertiary level. In Hyland (2004, p. 187) and Ferdous (2013, p. 44), academic texts or, more specifically, academic articles were even the most preferred genre.

Only few studies have documented little reading time (Avello et al., 2019; Bengtson, 2014; Kaur, 2015), with Freeman (1999, p. 85) being the first to thematize a lack of reading despite the availability of material and sources. Similarly, Bengtson (2014, p. 19) states that the 42 Japanese language users involved in his study spent only an average time of 3.93 hours per week on English reading practices. This is even undercut by Knight's (2007, p. 52) study of exchange students located in Portland, U.S., where only 88.8 minutes per week were invested on reading. Kaur's (2015) qualitative study examining the EE habits of Malaysian tertiary level students sheds some light on possible reasons for reluctant use of reading. Apart from lack of availability of English sources, participants mentioned that they had too little time to read and that the texts found were often perceived as too difficult and therefore strenuous to read (Kaur, 2015, p. 7). Other studies (e.g., Arnbjörnsdóttir, 2017; Busby, 2020; Trinder, 2015) suggest that instead of an overall decline of reading literacy, habits have shifted from print media to online media, similar to reading practices at secondary school level. Trinder (2015, p. 2), for example, reported that 45% of her Austrian participants involved in tertiary education spent time on reading English online news portals and journals daily. Likewise, Arnbjörnsdóttir's (2017, p. 46) qualitative investigation of EE practices with the help of language diaries found that most reading was done online by the 32 Icelandic participants involved. More directly, Busby (2020, p. 15) summarizes that online reading is very popular amongst tertiary level students while consumption of print media varies considerably.

Similar to the observed increase of reading practices at tertiary level, the time spent on speaking activities has become more varied. While at secondary school level participants were only very rarely involved in EE speaking activities (e.g., Lamb, 2004a; Lamb, 2004b; Sundqvist, 2009), several studies at tertiary level (Freeman, 1999; Knight, 2007; Lamb, 2002; Moncrief, 2011; Spratt, Humphreys & Chan, 2002) suggest increased communication outside of educational settings. In her study of in-work and out-of-work EE practices of participants in Portland of varying proficiency levels, Knight (2007, p. 52) arrives at a mean time of 361.3 min/week for work-related speech and 104.06 min/week spent on

conversations in English with friends. This amounts to approximately 7.76 hours per week spent on speaking English. However, one needs to consider that the study is located in an English-speaking environment, increasing the necessity to use the English language for communication. Yet, other studies based in an EFL environment (e.g., Broemer & Virkkunen-Fullenwider, 2011; Ferdous, 2013; Hyland, 2004; Kaur, 2015; Kusyk, 2017; Liu, Lin & Wiley, 2016; Pearson, 2004; Pickard, 1996; Sockett & Toffoli, 2012) indicate that EE speaking activities have a similar status at tertiary level as on secondary school level, with participants speaking English rather scarcely outside of an educational setting. While participants in Pickard (1996, p. 4) mention limited opportunities to speak English in their spare time as reason for their limited investment, students in Kaur (2015, p. 7) additionally admit a lack of initiative. Furthermore, Lamb's (2002, p. 43) qualitative study of 16 Indonesian university students highlights how cultural factors can interplay with EE engagement, since his participants stated that they avoided speaking English outside of the classroom due to social norms. During the interviews, students reported that they were being made fun of when speaking English outside class and that they were considered arrogant when speaking English to Indonesian people (Lamb, 2002, p. 43). Hence, cultural and national background might influence participants choice of EE practices.

Another difference detected when comparing studies on the tertiary level to research conducted in secondary schools is the focus on English computer games. While studies at earlier educational levels show a strong interest in the influence of gaming on second language proficiency, fewer studies appear to be concerned with this topic at tertiary level (e.g., Bengtson, 2014; Busby, 2020; Ferdous, 2013; Liu, 2014; Lai, Hu & Lyu, 2018; Moncrief, 2011; Orhon, 2018). While Lai, Hu and Lyu (2018, p. 120) report that tertiary students are indeed involved in playing English computer games, studies overall do not agree on the importance of gaming as extramural activity influencing language learning at tertiary level. While Orhon's (2018, p. 7) findings on EE activities suggest that 108 Turkish university students are indeed involved in playing second language computer games regularly, Moncrief (2011, p. 112) and Ferdous (2013, p. 51) attest a weak standing of gaming in their studies of Finnish and Asian students. However, while in Moncrief's (2011, p. 112) study gaming is placed last with only one student engaging in active English gaming, playing computer games is more popular in Ferdous' (2013, p. 51) study with at least 15% playing very often. Other studies, such as Bengtson (2014, p. 23), Liu (2014, p. 77) or Busby (2020, p. 15) display mixed findings with great variance regarding the involvement in extramural gaming practices at tertiary level.

Yet, comparison of educational levels does not only reveal differences but additionally allows the identification of similarities. In line with studies on primary and secondary level (e.g., De Wilde & Eyckmans, 2017; Lefever, 2010; Kuppens, 2010; Persson & Prins, 2012), listening to English music (e.g., Broemer & Virkkunen-Fullenwider, 2011; Ekşi & Aydin, 2013; Ferdous, 2013; González-Fernández, 2021; Inozu, Sahinkarakas & Yumru, 2010; Knight,

2007; Kusiak, 2017; Littlewood & Liu, 1996; Pickard, 1996; Liu, 2014; Mukundan, Khojasteh & Pearson, 2009; Orhon, 2018) and watching video material in English (e.g., Avello et al., 2019; Broemer & Virkkunen-Fullenwider, 2011; Ekşi & Aydın, 2013; Ferdous, 2013; González-Fernández, 2021; Littlewood & Liu, 1996; Liu, Lin & Wiley, 2016; Knight, 2007; Moncrief, 2011; Orhon, 2018; Pearson, 2004; Sockett & Toffoli, 2012; Trinder, 2015) are very popular extramural activities especially in more recent research. However, these activities appear not to be as dominant at tertiary level as at other educational stages. Moreover, in agreement with studies at the secondary school level, several studies (e.g., Inozu, Sahinkarakas & Yumru, 2010; Kaur, 2015; Moncrief, 2011; Mukundan, Khojasteh & Pearson, 2009; Trinder, 2015) report that the use of the internet is a popular EE activity at tertiary level. In contrast to other studies, Trinder (2015, p. 405) does not treat the internet as a holistic concept but differentiates between consultation of informational websites, visiting company websites, and engaging in social media. While the use of informational websites, such as Wikipedia, is the most popular online activity with 71% using them frequently, participation in social media activities is comparable with 58% using social media in English daily (Trinder, 2015, p. 405). Similar results concerning the use of social media were identified by other studies (e.g., Ferdous, 2013; Kusiak, 2017; Sockett, 2013; Sockett & Toffoli, 2012), with social media being the most popular productive EE activity in Kusiak's (2017) study of 943 French and German university students.

Regarding the engagement in receptive and productive extramural activities, studies at tertiary level largely correspond to findings at other educational stages, with participants voicing a clear preference for receptive EE activities in comparison to productive involvement (e.g., Arnbjörnsdóttir, 2017; Chusanachoti, 2009; Hyland, 2004; Kusiak, 2017; Lai & Zheng, 2018; Littlewood & Liu, 1996; Orhon, 2018; Pickard, 1996). As Pickard (1996, p. 152) has already observed early on, “[t]he majority of the data collected relate[s] to the receptive skills of reading and listening; data on speaking and writing [is] negligible”. Especially for EE activities concerned with writing, a persistent lack can be determined through all educational levels. The little involvement in EE writing activities found amongst tertiary level students (e.g., Ekşi & Aydın, 2013; Ferdous, 2013; Kaur, 2015; Knight, 2007; Orhon, 2018), strengthens the assumption brought forward by Kaur (2015) that participants see the classroom as the main domain for written second language activities. Nevertheless, participants appear to engage in some EE activities related to writing, since Kusiak (2017, p. 83) reports that writing e-mails was the most used productive activity in her study (1-3 hours a month). Similarly, Arnbjörnsdóttir (2017, p. 49) reports that her participants engaged in writing of English blogs and social media messages.

Taking a look at the EE activities discussed so far, one can subsume them under the label of informal EE activities, being actions that mainly do not involve conscious L2 learning. While formal extramural activities were negligible at prior educational levels, increased autonomy at the tertiary level appears to foster engagement in conscious extramural language learning

practices (e.g., Ekşi & Aydın, 2013; Inozu, Sahinkarakas & Yumru, 2010; Lai, Hu & Lyu, 2018; Lamb, 2002; Littlewood & Liu, 1996; Liu, 2014; Pearson, 2004; Pickard, 1996; Shen et al., 2005). Amongst these more formal extramural activities, participants engaged in independent study in the library (Pearson, 2004), using self-study material (Lai, Hu & Lyu, 2018; Shen et al., 2005), seeking help from the teacher (Shen et al., 2005), doing vocabulary and grammar exercises (Inozu, Sahinkarakas & Yumru, 2010; Liu, 2014), studying new words (Ekşi & Aydın, 2013) and consulting a dictionary (Ekşi & Aydın, 2013; Inozu, Sahinkarakas & Yumru, 2010; Mukundan, Khojasteh & Pearson, 2009). While some studies (Chan, Humphreys & Spratt, 2002; Mukundan, Khojasteh & Pearson, 2009; Spratt, Humphreys & Chan, 2002; Trinder, 2017) do not agree with the prominence of formal EE activities at the tertiary level, there is consent on the increase of dictionary use for EE activities.

Regarding the influence of EE activities on other variables, similar findings can be reported at tertiary level as for other educational levels. Knight (2007, p. 59) detected a significant positive correlation of quantity of EE activities and scores on the PPVT ($r=.318$; $p= 0.021$). Similar results were reported by Kalaja et al. (2011), Miller and Hegelheimer (2006) and Cole and Vanderplank (2016). However, the use of self-designed tests, which are not described in much detail, by Miller and Hegelheimer (2006) and Cole and Vanderplank (2016) makes it difficult to judge whether significant results relate to receptive or productive vocabulary knowledge. Nevertheless, Busby (2020, p. 16) has identified a similar positive relationship as Knight (2007, p. 59) for receptive vocabulary knowledge measured with the help of the VLT (Schmitt, Schmitt & Clapham, 2001) and “the frequency of reading books, gaming, and reading online in English”. In addition, a recent study by González-Fernández (2021) investigates the influence of 5 types of EE exposure, namely reading, watching TV, using social networks, playing computer games and listening to music, on the receptive and productive ability of four word knowledge components. While reading, watching TV and using social networks show positive correlations to all measures of vocabulary, listening to music is only linked to receptive vocabulary size measured by the VLT (González-Fernández, 2021, p. 36). When going more into depth with multiple linear regression models, González-Fernández (2021, p. 37) reveals that involvement in reading and watching TV function as significant predictors for receptive and productive form-meaning, collocation and derivational knowledge, explaining between 22% (derivational recognition knowledge) and 7% (collocation recognition knowledge). In comparison, receptive and productive multiple meaning knowledge is only predicted by involvement in reading, which explains 6% of receptive and 11% of productive multiple meaning knowledge (González-Fernández, 2021, p. 37). Without denying the significant insights of this study, it has to be noted that González-Fernández (2021) included a broad range of participants aged between 18 and 65 years of various proficiency levels, who were tested on high-frequency words, displaying a wide spectrum. Hence, it will be seen whether similar findings can be reported for the narrower scope of academic verb performance of English major students in this study.

In contrast to the findings reported above, Lee (2019a, p. 767 & 2019b, p. 120) could identify no significant relationship (PVLT $r = 0.04$, $p > 0.05$, VLT $r = 0.09$, $p > 0.05$; PVLT $r = 0.13$, $p > 0.05$) between quantity of EE activities and productive and receptive vocabulary knowledge. Instead, he reports that diversity of extramural practices influenced the scores on both vocabulary tests significantly (PVLT $r = 0.46$, $p < 0.01$, VLT $r = 0.27$, $p < 0.05$; PVLT $r = 0.54$, $p < 0.01$) (Lee, 2019a, p. 767; Lee 2019b, p. 120). Comparable results regarding a potential relationship between productive vocabulary knowledge and diversity of EE use ($r = 0.51$, $p < 0.01$) are also reported by Lee and Dressman (2018, p. 440). Hence, more research will have to be conducted to evaluate whether both, quantity and quality of EE, are related to vocabulary knowledge and whether one factor exerts a stronger influence.

Another language ability researched in connection to EE is speaking, with studies showing a positive relationship between quantity of EE and communication skills (Reinders & Wattana, 2012; Suh et al., 1999). In their interview study of 8 international students, Suh et al. (1999, p. 11) students felt that they were able to improve their speaking skills by engaging in EE activities. Moreover, Reinders and Wattana (2012) observed a positive relationship between the quantity of EE practices and participants' willingness to communicate in the language. However, Lee (2019b, p. 120) did not find a significant relationship between the amount of EE and speaking ability ($r = .22$, $p > 0.05$). Instead, Lee (2019b, p. 120) and Lee and Dressman (2018, p. 440) report that, similar to vocabulary knowledge, diversity of EE had a positive influence on speaking performance ($r = 0.56$, $p < 0.01$; $r = 0.56$, $p < 0.01$).

While other language skills, such as grammatical knowledge (Cole & Vanderplank, 2016) or writing ability, still appear to be underrepresented in the research field, more interest towards overall language proficiency and grades is visible. As the investigations of Lee (2019b, p. 120) and Littlewood and Liu (1996) suggest, no clear relationship between overall language proficiency, tested via different proficiency tests, and amount of EE can be found. While Lee's (2019b) study arrives at no statistical relationship between the TOEIC proficiency test and quantity of EE ($r = 0.24$, $p > 0.05$), Littlewood and Liu (1996) found a positive relationship for one of their two tests administered but not for the other. However, as for vocabulary knowledge and speaking skills, the diversity of EE activities appears to be an influential factor (Lee, 2019b). Similarly, ambiguous results are found when comparing students' self-evaluation of overall language proficiency to amount of EE activities. In contrast to Arnbjörnsdóttir (2017, p. 49-50), who was able to detect a descriptively positive relationship between the two variables, Lai, Hu and Lyu (2018, p. 129) could not identify a statistically significant link. At the same time, Shen et al. (2005) were able to detect a positive relation between amount of EE exposure and scores on a university entrance exam and overall English grades. However, the ambiguity of results might partially be explained by the fact that the concepts tested, namely language proficiency and grades, are very latent and broad concepts, which allow for much variation of interpretation.

5.3 Challenges of EE and possible solutions

Moving one step further from examining EE practices and their link to language learning triggers questions relating to EE's actual and desirable effects on learning in the classroom. Yet, pedagogical implications of findings for the educational context need to be made with caution, since EE is defined by its detachment from any educational context and figures. Consequently, the consideration of this area of tension appears to be crucial for every research study aiming at the formulation of pedagogical implications.

Despite the positive effects of EE on various aspects of language learning, researchers have described problems and challenges caused by out-of-class English language contact. Due to the rapid change of the English media landscape, especially in Scandinavian countries (Cabau, 2009, p. 134), teachers are faced with new obstacles (Henry et al., 2018). As several researchers (e.g., Henry et al., 2018; Olsson, 2011; Sundqvist & Olin-Scheller, 2013) observed, teachers are struggling to bridge the gap between informal out-of-class English contact and in-class English activities. Students often bemoan the insufficiency of authenticity of classroom activities in comparison to their spare time EE activities (Henry et al., 2018, p. 251), resulting in a lack of motivation in class (Mukundan, Khojasteh & Pearson, 2009). In the following, two approaches to tackle this newly evolved phenomenon are presented:

Keeping this tension in mind, several studies (e.g., Bunting & Lindström 2013; Grau, 2009; Henry et al., 2018; Ingvarsdóttir & Jóhannsdóttir, 2018; Toffoli & Sockett, 2016) focused on the investigation of teachers' awareness and integration of students' EE activities into their teaching. A large study by Ingvarsdóttir and Jóhannsdóttir (2018) examined the perception of 703 Icelandic pupils on the use of English inside and outside of school both in the present and in the future. Results show that EE activities of students are mostly not noticed or integrated into the classroom (Ingvarsdóttir & Jóhannsdóttir, 2018). Instead, teachers stick to traditional teaching methods which do not reflect students' use of English in their free time. In a German context, Grau (2009) found that teachers were only partially aware of their students' EE activities. Consequently, English outside and inside of school were treated as "two separate spheres" (Grau, 2009, p. 160). Similar results were reported by Bunting and Lindström (2013, p. 207) in their study of 47 11-year-old Swedish students. During pairwise interviews, they found that students clearly differentiated between accepted and non-accepted learning practices in class (Bunting & Lindström, 2013, p. 211). To be more precise, while some EE activities were more integrated into the classroom, such as watching TV, other activities as playing computer games were reserved for their spare time. This corresponds to Lai, Yeung & Hus' (2016) discovery that teachers had a significant influence on students' perception of how valuable specific EE activities were. Hence, teachers should be aware and mindful with their input on technological out-of-class activities. Wilkinson (2002) addresses this problem of teacher influence on EE activities in his ethnographic study

of 4 non-native speakers of French spending a month abroad. He discovered that some native speakers adopted a classical teacher role in the communication with their non-native host, which was considered positive by most of the non-native speakers (Wilkinson, 2002, p. 160). Conversely, non-native speakers also tried to introduce classroom norms into the conversation themselves (Wilkinson, 2002, p. 168). Hence, the instructional classroom setting impeded heavily on communication in an EE setting (Wilkinson, 2002, p. 168).

Conversely to the findings of the studies above, Henry et al.'s (2018) study of 97 Swedish EFL teachers, Schurz et al.'s (2022) study on 20 EFL teachers from 3 countries and Helgeson's (2018, p. 13) master thesis reveal that teachers are generally positively aware of their students' EE practices. Similar findings are also reported by Bergström, Norberg and Nordlund (2021, p. 10), who report that the 14 Swedish EFL teachers interviewed showcase a strong belief in incidental vocabulary learning and perceive vocabulary as being "picked up along the way". Schurz et al.'s (2022) analysis of 20 teacher interviews in the contexts of Sweden, Austria and France paints a similar picture. Teachers across all three countries believed that EE impacts fluency and language learning overall. However, slight differences were found regarding teachers' beliefs in the impact of EE on grammar knowledge.

In the context of tertiary education, Liu (2014) and Toffoli and Sockett (2016) investigated the perception and integration of EE activities into teaching at university level in France and China. Both Liu's (2014) quantitative study of 160 university students and Toffoli and Sockett's (2016) qualitative study of 30 university professors indicated that lecturers were aware of EE activities that their students were involved in. However, while most of the university teachers thought that EE activities had a positive impact on their classroom, only few of them tried to incorporate them into their teaching (Toffoli & Sockett, 2016). Some potential reasons for this reluctance reported were the fear of illegal practices online, avoidance of intruding into students' private life, the perceived poor quality of English consumed and insufficient technical knowledge on the side of the teacher (Toffoli & Sockett, 2016). Grau (2009) lists additional reasons for the separation of in-class and out-of-class language activities mentioned during her focus group interviews. While teachers admitted "scepticism regarding students' language competence" (Grau, 2009, p. 168), students feared that teachers might invade their private life (p. 169) and that the generation gap between teacher and students was too large (p. 170). Additionally, both teachers and students commented on the felt difference between classroom English and English used in their free time as limiting factor (p. 170). Helgeson (2018) found similar reasons for the avoidance of new media in classroom activities as Toffoli and Sockett (2016). When being asked why they did not incorporate new media more into their classroom activities, 73% of all teachers responded that they lacked the knowledge how to integrate new media into their English teaching. Other reasons named were that the technical equipment was not sufficient (36%) or that teachers felt they did not have enough time (36%) (Helgeson, 2018, p. 18). On the other side, reasons for the use of new media in the classroom were that teachers felt that

they could catch pupils' attention more easily and because students asked them to do so or because the teachers themselves were interested in EE activities (Helgeson, 2018, p. 19).

This corresponds to Henry et al.'s (2018) findings in their examination of how teachers try to bridge the gap between in-class and out-of-class English activities. By asking 97 teachers to describe an in-class activity that they felt had motivated their students, Henry et al.'s (2018, p. 256) study could show that the integration of popular culture was a common strategy of teachers, including the interpretation and comparisons of popular films to the original book or an analysis of music videos and blog posts. Moreover, teachers mentioned that tasks fostering contact and identification with the target culture increased students' interest (Henry et al., 2018, p. 256). Students reacted positively to such activities; for example when in-class tasks were related to their everyday life, future plans and desires (Henry et al., 2018, p. 257) or when they were allowed to choose the topic (Henry et al., 2018, p. 262). Furthermore, student motivation seemed to be increased by challenges in relation to such activities, such as competitive games or activities waking curiosity like watching only the beginning of a movie (Henry et al., 2018, p. 260-261).

Regarding the source of materials, teachers reported a more positive reaction to authentic teaching material, followed by authentic material adapted by the teacher or no material at all (Henry et al., 2018, p.258-259). This relates to students' increased motivation through creative activities in class (Henry et al., 2018, p. 261). However, concerning output, teachers mentioned that both task types, including a final authentic product, such as a radio programme, and tasks which required students to produce non-authentic products connected to the school context, such as a presentation or a portfolio, were successful (Henry et al., 2018, p. 259-260).

As a potential solution, Sundqvist and Olin-Scheller (2013) and Liu (2014) propose that further training in the area of EE should be offered. This is backed up by a study of fourteen EFL teachers taking part in a study programme called BOOST, which should help them to bridge the gap between out-of-class and in-class English activities (Sundqvist & Olin Scheller, 2013). In the survey, most teachers reported that they adapted their methodology after the course towards more integration of new media into the classroom. Moreover, all but one participant stated that they felt more confident after taking part in the study programme, although integrating new media into the classroom meant a switch of power from the teachers as the sole expert to more student empowerment. Another example of computer-related teacher training is introduced by Sercu and Peters (2002, p. 32), who investigated the opinions of course developers and participating teacher trainees on the effectiveness of a self-tuition course on computer-assisted language learning. While the course certainly is a valuable step into the right direction, Sercu and Peters (2002, p.41) outline that the course design and content needs to be further adapted, since learners felt that the course content

dealt too much with basic computer skills, while asking them to complete too complex tasks.

While the studies presented above conceptualize the differences between in-class and out-of-class language activities as problematic and see a solution in the assimilation of in-class activities to students spare time language activities, other studies see complementary potential in the two spheres.

In her case study of the 12-year-old Simo from Finland, Koivistoinen (2015) argues that formal and informal settings should not be viewed as opposing fields but that language learning is part of everyday life and that teachers need to reconsider where the two spheres, inside and outside of school, might overlap and interact. Likewise, Cabot (2016) presents the relationship between in-school and out-of-school learning in a more positive light by stating that the two learning contexts are working largely complementarily. In his qualitative study of six students in their first year of secondary school in Norway, it crystalized that out-of-school learning and in-school learning fulfilled partly different functions (p. 169). While out-of-school learning offered many output opportunities, functions such as, “noticing, testing and metalinguistic functions of written output were scarce in all interviewees’ out-of-school learning, but seemingly well-balanced at school” (Cabot, 2016, p. 179). Similar to Koivistoinen’s (2015) and Cabot’s (2016) participants, Lai’s (2016) 11 undergraduate interview partners used their resources in and outside of school in a complementary manner (Lai, 2016, p. 265). They reported that the classroom was the setting where basics of the language, such as grammar, vocabulary, or other language skills, are acquired and trained (Lai, 2016, p. 272). Moreover, they stated that the classroom helped them to be persistent, since they felt they would not have enough motivation to learn the language without teacher support (Lai, 2016, p. 272). While in-class learning fulfilled the function of being externally promoted by the teacher, out-of-class learning served several different functions, such as triggering the interest in learning the language, using the language for enjoyment and self-assessment through testing ones’ abilities (Lai, 2016, p. 272). Additionally, students mentioned intrinsic differences between the two spheres of language learning, such as in-class learning being compulsory and timed and out-of-class language learning being voluntary and spontaneous (Lai, 2016, p. 273). Moreover, students highlighted that out-of-class, they did not intend to learn but that they were focused on the activity itself which they did for enjoyment (Lai, 2016, p. 273). Overall, the pupils felt that in-class and out-of-class learning both served relevant functions and that the two spheres could be used successfully in a complementary manner (Lai, 2016, p. 275). For instance, in-class experiences or tasks encouraged them to expand on certain topics in their spare time and helped them to focus their interests and needs (Lai, 2016, p. 275). Conversely, out-of-class learning helped them to remember why they are learning the language, supported enjoyment of the language in and outside the classroom and made the language classroom more authentic (Lai, 2016, p. 276).

In their study of 2015, Lai, Zuh and Gong investigated the different functions of out-of-class learning in relation to in-class language learning by conducting focus group interviews with 19 middle school students in China. While students described their in-class learning as form-focused, knowledge focused and rather teacher-centred (p. 288), their out-of-school language contact was meaning-driven and content-oriented (p. 292). This complementary nature of out-of-school learning was predictive for good learning outcomes (p. 298). Hence, out-of-school language contact is crucial as a complementing sphere to in-school teaching. Similar to Lai, Zuh & Gong (2015), Sockett (2013, p. 52) asserts that out-of-school language contact is mainly driven by meaning instead of form and has a focus on enjoyment of communication. By analysing blog posts of 9 female university students about their extramural language activities, he identified key features of EE activities (Sockett, 2013, p. 51). Firstly, learning mostly occurs through discussion and negotiation with a conversation partner or themselves (Sockett, 2013, p. 52). Additionally, learners make use of acquired strategies such as skimming of texts to quickly find specific information in social media posts (Sockett, 2013, p. 53). Moreover, cognitive tasks such as learning new vocabulary items are aided by meaning-centred goals, such as the understanding of a specific text (Sockett, 2013, p. 54). Therefore, it is of crucial importance that students choose texts which are of intrinsic interest to them (Sockett, 2013, p. 55). Another positive function of out-of-school language contact found in the data was that students detected language patterns, such as the existence of specific recurring grammatical structures or the existence of different registers, leading to raised awareness for language structures and finally imitation thereof (Sockett, 2013, p. 55-56). This shows that noticing can be both teacher (Cabot, 2016) and extramurally initiated (Sockett, 2013).

A more practice-oriented study by Jakonen (2014), analysing 30 conversational episodes recorded in CLIL history lessons, highlights opportunities where extramurally acquired knowledge can be applied by students. He observed that while whole-class interactions seemed to offer less opportunities for students to bring in extramural language knowledge, peer group interaction were often used by students to use and refer to extramural language sources (Jakonen, 2014, p. 22).

The comparison of the studies outlined reveals crucial functions of in-class and out-of-class language activities for language learners (see Table 5.2).

Functions of out-of-school activities	Functions of in-school activities
offer output opportunities	covers noticing, testing and metalinguistic functions
increase interest and enjoyment in language learning	train basic language skills
enhance self-assessment	aid persistence

mainly implicit learning	mainly explicit learning
are meaning-driven and content-oriented	are form-driven and knowledge-focused
support learning through negotiation	can be more teacher-centred
build on acquired skills, such as skimming	
initiate detection and imitation of language patterns	

Table 5.2: Complementary functions of in- and out-of-school language learning

In a next step, more studies in this area need to be conducted to create a picture of the functions of the two spheres for language learning that is as complete as possible. This will hopefully allow us to make assumptions of how in-class learning really has to be altered to adequately respond to EE activities of students, instead of precipitously jumping to the conclusion that the integration of extramural new media activities into the classroom will be the best solution.

6 Study design

Building on the insights gained from the literature review this study contributes to two research fields, namely research on the nature of vocabulary knowledge and Extramural English, as well as their potential links. More precisely this research study intends to address the following research gaps identified:

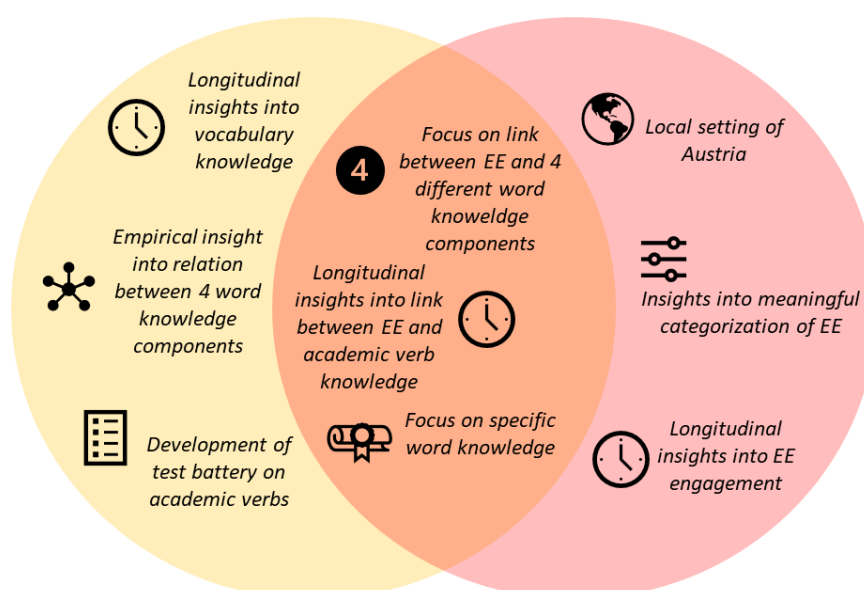


Figure 6.1: Main research gaps addressed

The research gaps related to the field of Extramural English are visualized in a red circle, new insights related to the field of vocabulary knowledge in a yellow circle and in the intersection new findings linked to both areas of investigation are listed.

Starting with gaps connected to the field of EE, this study strives to enlarge the relatively small pool of EE research conducted in the geographical context of Austria, allowing insights into the EE engagement of Austrian tertiary students. Due to the study's longitudinal design it additionally hopes to offer rare insights into the consistency of EE engagement over the course of one academic year, since prior studies have been mostly cross-sectional or pseudo-longitudinal studies. Furthermore, prior EE studies have often grouped EE activities based on theoretical conceptualization, which has been complemented by statistical means of meaningful categorization in this study.

Moving on to novel insights provided for the field of vocabulary knowledge this study offers empirical insights into the relation between four different word knowledge components. For this purpose a test battery including one test for each word knowledge component to be examined had to be developed for the target vocabulary of academic verbs. Although also prior studies already pursued to examine the relationship between two (e.g. Zhang & Lu, 2015; Zhong, 2018) or sometimes also more (González-Fernández & Schmitt, 2019) word knowledge components, most studies focused on a different type of vocabulary knowledge. Moreover, this study moves beyond classical correlational analysis and strives to investigate the suitability of a model inspired by Nation's (2013) taxonomy of word knowledge. This is further complemented by the longitudinal nature of this research, which allows an observation of the stability of relationships and models over time.

Lastly, while a large number of studies (e.g. Schwarz, 2020) has focused on the relationship between EE and general vocabulary knowledge, mostly defined as vocabulary size measured through receptive and/or productive form-meaning link knowledge, few studies (González-Fernández, 2021) have focused on multiple word knowledge aspects and their relation to EE engagement. Moreover, few prior studies (Busby, 2020; Olsson & Sylvén, 2015; Sundqvist, 2009; Warnby, 2022) have laid their focus on specific word knowledge, such as academic word knowledge. Consequently, this study aims to investigate whether EE contact can also be a potential predictor for specific word knowledge, more precisely for academic verb knowledge. Furthermore, this study hopes to offer rare longitudinal insights into the stability of potential predictor relations and their strength.

6.1 Research questions

Based on the research gaps outlined above the following research questions have been formulated for this study. To increase clarity research questions are assigned to one of the research areas investigated as visualized in Figure 6.1. Therefore, the section first lists all research questions concerned with the field of Extramural English, followed by a section on

insights related to vocabulary knowledge, succeeded by a section on research questions related to both fields. For each of the areas examined overarching research questions are given first, followed by more detailed sub-questions.¹⁰ Research questions answering one of the research gaps visualized in Figure 6.1 are assigned the same symbol as the gap in the figure to allow easier matching of research questions with the gaps addressed.

A) Extramural English

 *What kinds of extramural English contacts do English major students in Austria report to have?*

 *Does this exposure to EE change over the course of one academic year?*

 *How could this EE contact be categorized in a meaningful way based on theoretical and statistical considerations?*

- What kinds of recording techniques are used by English major students in Austria in EE settings?
- Do these recording techniques change over the course of one academic year?
- Which word knowledge components are considered by English major students in Austria in EE settings?
- Does the attention paid to different word knowledge components in EE settings change over the course of one academic year?
- Is there a connection between biographical background variables and participants' exposure to EE, their recording techniques used and the word knowledge aspects considered in EE settings?
- Which biographical background variables show a connection to EE exposure, recording techniques and word knowledge aspects considered over the course of one academic year and which variables are not stable?

B) Vocabulary knowledge

 *Which development of academic verb knowledge is observable over the course of 1 academic year?*

- Which development of form-meaning recognition knowledge of academic verbs is observable over the course of one academic year?
- Which development of form-meaning recall knowledge of academic verbs is observable over the course of one academic year?
- Which development of collocation recognition knowledge of academic verbs is observable over the course of one academic semester?
- Which development of multiple meaning recognition knowledge of academic verbs is observable over the course of one academic semester?

¹⁰ For each research question the formulated H_0 and H_A can be found in appendix 6.1 and/or it has been noted that the question has been answered descriptively. Validated hypotheses are highlighted in bold.

-  Which relationship can be found for the four word knowledge components observed?
-  How stable are relationships between the four word knowledge components observed with the help of correlational analysis and a SEM model?

C) Relationship between EE and academic verb knowledge

 Is there a relationship between EE activities and academic verb knowledge?

- Is there a relationship between EE activities and form-meaning recognition knowledge of academic verbs?
- Is there a relationship between EE activities and form-meaning recall knowledge of academic verbs?
- Is there a relationship between EE activities and collocation recognition knowledge of academic verbs?
- Is there a relationship between EE activities and multiple meaning recognition knowledge of academic verbs?

4 Which EE activities can be identified as predictors of academic verb knowledge?

- Which EE activities predict form-meaning recognition knowledge of academic verbs?
- Which EE activities predict form-meaning recall knowledge of academic verbs?
- Which EE activities predict collocation recognition knowledge of academic verbs?
- Which EE activities predict multiple meaning recognition knowledge of academic verbs?

 How stable are the relationships between EE activities and the four different types of academic verb knowledge over one academic year?

D) Word knowledge components considered in EE settings & academic verb knowledge

Is there a relationship between types of word knowledge components considered in EE settings and academic verb knowledge?

Which types of word knowledge components considered in EE settings influence academic verb knowledge?

- Which types of word knowledge components considered in EE settings influence form-meaning recognition knowledge of academic verbs?
- Which types of word knowledge components considered in EE settings influence form-meaning recall knowledge of academic verbs?
- Which types of word knowledge components considered in EE settings influence collocation recognition knowledge of academic verbs?
- Which types of word knowledge components considered in EE settings influence multiple meaning recognition knowledge of academic verbs?

E) Biographical variables & academic verb knowledge

Is there a relationship between biographical variables and academic verb knowledge?

Which biographical variables influence the development of academic verb knowledge?

- Which biographical background variables influence the development of form-meaning recognition knowledge of academic verbs?
- Which biographical background variables influence the development of form-meaning recall knowledge of academic verbs?
- Which biographical background variables influence the development of collocation recognition knowledge of academic verbs?
- Which biographical background variables influence the development of multiple meaning recognition knowledge of academic verbs?

6.2 Data collection & Instruments

In the following chapter a brief overview of the data collection instruments used to answer the above mentioned research questions will be given.¹¹ In addition, an overview of the data collection process and the data collected for this study will be given.

As already foreshadowed in the research questions listed above this study is a mainly quantitative longitudinal study encompassing data collected within one academic year.

6.2.1 Collection of data related to vocabulary knowledge:

To answer the research questions related to academic verb knowledge four vocabulary tests, tapping into form-meaning recognition, form-meaning recall, collocation recognition and multiple meaning recognition of academic verbs were developed¹². As visible in Figure 6.3 the Form-meaning Recognition Test of Academic Verbs and the Form-meaning Recall Test of Academic Verbs were developed first and were primarily piloted with a group of native speakers and university colleagues, followed by a second round of piloting in January with a group of students comparable to the target group, before they were used in the first data collection in March after analysing the data from the two piloting rounds. All data was collected before the pandemic and all collections were taking place anonymously¹³.

As visible in Figure 6.2 147 Form-meaning Recognition test booklets and 147 Form-meaning Recall test booklets were analysed for the 1st data collection. Therefore, 294 test booklets were scored and in a next step entered into the program SPSS (IBM, 2020). The scoring and

¹¹ While an extra chapter has been dedicated to the detailed description of test development, the EE questionnaire will be mainly described within this chapter.

¹² See chapter 7 for details on test development.

¹³ For more detail see chapter 6.3.

the data entry procedure was done by the researcher and a colleague¹⁴ who double checked all scores as well as the data entered into SPSS (IBM, 2020) between April and May.

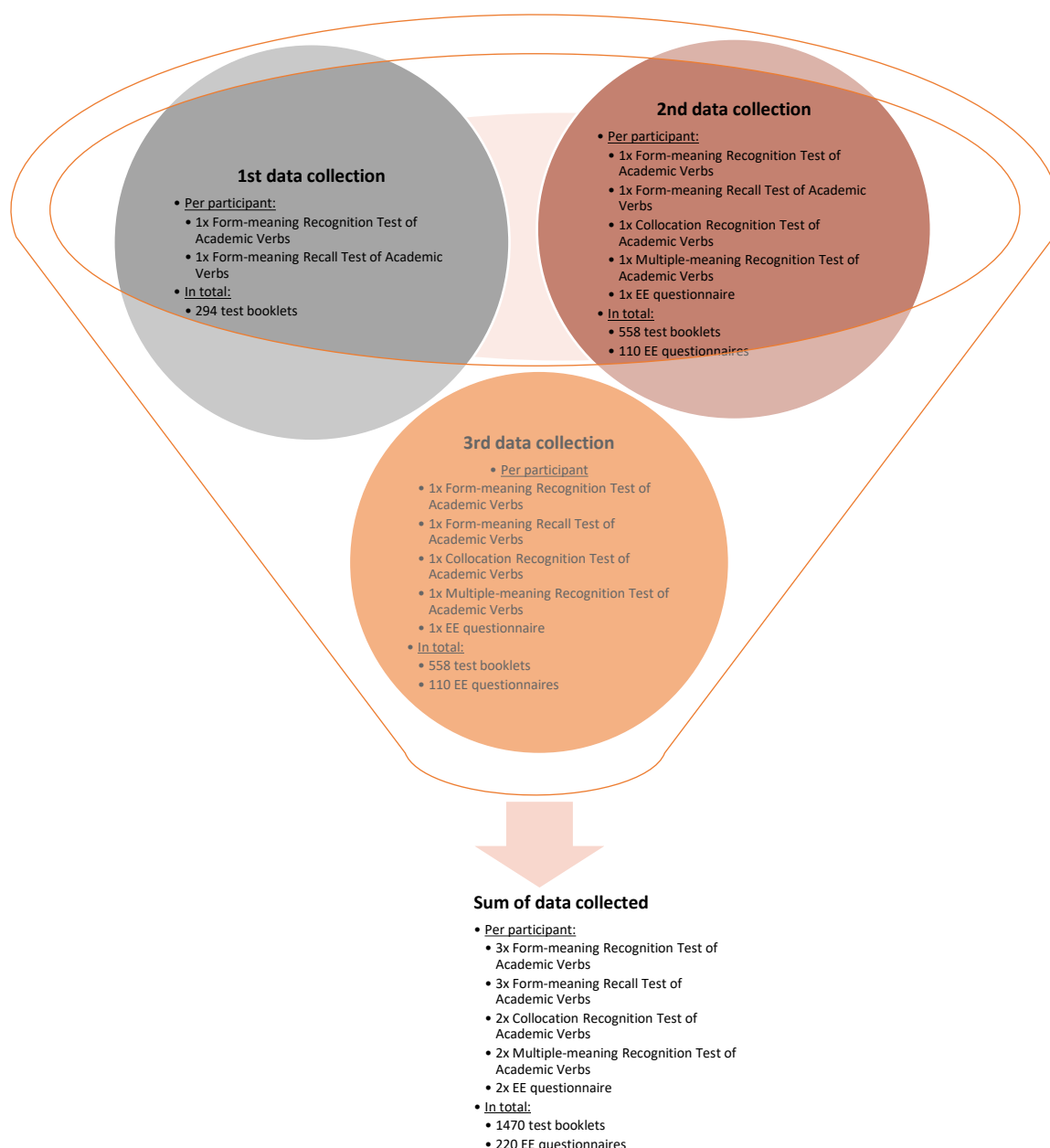


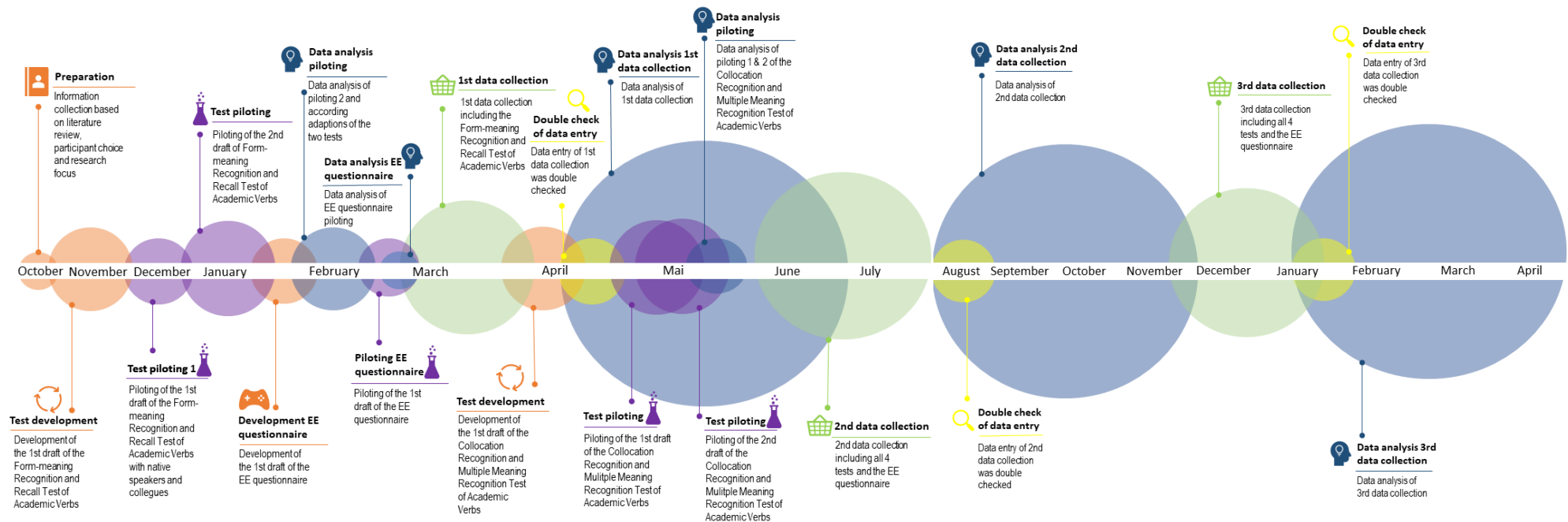
Figure 6.2: Sum of data collected and analysed

Concurrent to these steps the first drafts of the Collocation Recognition Test of Academic Verbs and the Multiple Meaning Recognition Test of Academic Verbs were developed, which were piloted twice in May, resulting in the final tests, which were then integrated into the second data collection in June and July.

¹⁴ Thanks go again to Mag. Magdalena Stockhammer, MA.

The second data collection resulted in another 147 Form-meaning Recognition, 147 Form-meaning Recall, 147 Collocation Recognition and 147 Multiple-meaning Recognition test booklets. This amounts to a sum of 558 test booklets, which were scored, entered and double checked for the second data collection.

In the third data collection taking place in December and January the same number of test booklets was collected as for the second data collection, which results in 1470 test booklets, which were collected, scored, entered, double checked and analysed for this study in total (see Figure 6.2)



Orange: Processes of instrument development

Violet: Piloting rounds of study instruments

Blue: Data analyses

Green: Main data collection points

Yellow: Double checks of data entry

Figure 6.3: Data collection process

6.2.2 The Extramural English questionnaire

To investigate participants' use of English outside the classroom, a questionnaire was developed simultaneously to the second piloting of the form-meaning link tests and was piloted at the beginning of March.¹⁵

The Extramural English questionnaire used in this study is based on Schwarz's (2020) EE questionnaire, which is grounded on a thorough analysis of other existent EE questionnaire, conglomerating various concepts and approaches. Nonetheless, several adaptations were made to compose a suitable questionnaire for the target group of this study. Due to Schwarz's (2020) participant choice of secondary school children the language used in the original questionnaire is the pupils' first language, namely German. Hence, the questionnaire had to be translated to English and additionally certain rubrics were deleted or shortened.

While Schwarz's (2020) EE questionnaire comprised 10 pages, the EE questionnaire in this study was reduced to 6 pages, following Dörnyei's (2007, p. 110) suggestion that most researches consider questionnaires longer than maximal 6 pages too long. Due to the different focus and target group used in this study no information about school grades or information about parents or closer family background was necessary. Moreover, this study refrains from focusing on the type of technical gadgets used to complete an EE task. Instead the focus is more on the modality of the EE activities used, such as audio-visual or auditory input and output. In addition, questions about EE engagement are overall sorted by language skill and not by the type of activity. Finally, the self-assessment section included in Schwarz's (2020) questionnaire for all language skills was excluded as well. Instead, questions were added about attention to new words in the spare time, recording techniques used for new words encountered in the spare time, as well as questions tapping into the consideration of different word knowledge aspects in the spare time.

In total the following sub-constructs consisting of multi-item scales as defined in Dörnyei (2007, p. 103-104) were included in the first draft of the EE questionnaire:

¹⁵ For the first and final draft of the EE questionnaire, as well as piloting results reported in this chapter see appendix 6.2, appendix 6.3 and appendix 6.4.

	Sub-constructs
Construct 1	Attention to new words (ANW)
Construct 2	English music (EM)
	Films in English (FIE)
	English online (EO)
	Reading in English (RIE)
	Writing in English (WIE)
	Speaking in English (SE)
	English theatre plays (ETP)
Construct 3	Conscious recording (CLW)
Construct 4	Word knowledge aspects (WA)

Table 6.1: Constructs of the EE questionnaire

It was decided to postpone questions regarding biographical information of the participants to the end of the questionnaire, since Dörnyei (2007, p. 111) and Oppenheim (1992) observed that personal qualification questions can be counteractive to the perception of anonymity of participation. Moreover, such classification questions usually resemble bureaucratic forms, which is rather counterproductive to catching participants interest (Dörnyei, 20007, p. 111). Therefore, based on the assumption that opening questions should be “interesting, relatively simple yet should at the same time [be] focused on important and salient aspects” (Dörnyei, 2007, p. 111), it was decided to start with questions regarding construct 1, which focuses on how much conscious attention students think they pay to new words encountered in their spare time. This topic was considered to be focusing on important and salient aspects related to the focus on vocabulary knowledge, which was also known to all participants. To ensure as accurate and comparable answers as possible students received the following information before starting to complete the questions for construct 1:

(21)

I am interested what you did in **English in your free time in the last semester outside of university**.

Ø Please only think of activities in your free time, which are not related to University (no assignments, research papers...).

To gather the information on attention to new words in the spare time a 5-point Likert scale, ranging from *(nearly) never* to *(nearly) always* as extreme ends, was utilized. Example 22 gives insights into how one of the items of construct 1 looks like.

(22)

How often do you pay attention to new words in your free time?	(nearly) never				(nearly) always
I pay attention to new English words if they are important for the content.	1	2	3	4	5

In the first draft construct 1 consisted of the following 5 items:

- I pay attention to new English words if they are important for the content.
- I consciously pay attention to new words in English media.
- As long as I understand what the content is about I am unobservant of new words.
- I do recognize new English words but I do not pay conscious attention to them.
- I am not interested in new words in my free time.

However, after the piloting of the first draft of the EE questionnaire at the end of March (see Figure 6.3) with 29 participants similar to the target group it was revealed that the reliability score of the construct could be increased from 0.64 to 0.69, if the third item was deleted.

	Sub-constructs	Cronbach's α before changes	Cronbach's α after changes
Construct 1	Attention to new words (ANW)	0.64	0.69
Construct 2	English music (EM)	0.75	0.74
	Films in English (FIE)	0.17	0.78
	English online (EO)	0.92	0.92
	Reading in English (RIE)	0.84	0.84
	Writing in English (WIE)	0.86	0.86
	Speaking in English (SE)	0.74	0.81
	English theatre plays (ETP)	0.60	X
Construct 3	Conscious recording (CLW)	0.49	0.49
Construct 4	Word knowledge aspects (WA)	0.84	0.84

Table 6.2: Reliability of constructs of EE questionnaire before and after piloting

Consequently, the final version of the EE questionnaire comprises only four items related to attention paid to new words.

The second construct included in the questionnaire concerns the use of English in the free time. Again participants are asked to focus on spare time only, excluding activities related to university, such as assignments or research papers to ensure selective answers. Similar to the items in construct one a 5-point Likert scale was used ranging from *(nearly) never* to *(nearly) always*. Overall, the following aspects of EE engagement were focused on in the first draft of the EE questionnaire:

- Engagement related to English music
- Engagement related to English films
- Engagement related to English use online
- Engagement related to English reading activities
- Engagement related to English writing activities
- Engagement related to English speaking activities
- Engagement related to English theatre plays

In more detail the first draft of the EE questionnaire comprised seven questions related to English music:

- Listen to English music in English
- Watching English music videos on the internet
- Singing English songs (karaoke etc.)
- Listening to English songs at a concert
- Reading English lyrics
- Translating English lyrics
- Writing English lyrics

Yet, the analysis of the piloting of the questionnaire in March showed that the reliability score of the scale could be increased from 0.74 to 0.75 (see Table 6.2), if item 6 (*Translating English lyrics*) was deleted. This was implemented in the final draft of the questionnaire, which left 6 items for the multi-item scale related to English music.

Regarding items related to watching English video material the first draft included items with and without the use of subtitles as can be seen in example 23 below.

(23)

Films in English					
Watch films in English with subtitles	1	2	3	4	5
Watch films in English without subtitles	1	2	3	4	5

However, the piloting revealed that reliability could be considerably increased from 0.17 to 0.78 (see Table 6.2), if the distinction between with and without subtitles was omitted. Therefore, the final EE questionnaire does not differentiate between consuming video

material with or without subtitles anymore. Instead the following specific open-ended question (Dörnyei, 2007, p. 107) was included to gain information on the use of subtitles:

(24)

⇒ How often do you use subtitles (Give percentage)? In which language(s) are the subtitles most of the time?
How often do you use which language? (*Leave out this question if you do not use subtitles.*)



The video material asked for in the final EE questionnaire encompasses watching English films, English series, English YouTube videos and English documentaries. In the first draft of the questionnaire participants were also asked to indicate whether they made YouTube videos themselves but similar to the subtitling questions this item had to be deleted for the sake of increased reliability.

Engagement with English online is concerned with using search engines in English, using English apps and playing single- or multiplayer games. While the reliability coefficient for this construct was very high with 0.92 in the piloting (see Table 6.2), some adaptations were necessary for the final EE questionnaire. Whereas the English online section was constructed as in example 25 below in the first draft of the questionnaire, it was discovered that participants in the piloting did not answer the questions about using search engines or English apps anymore because they thought they were connected to the statement in italics.

(25)

English online					
Playing games on the computer or on gaming consoles	1	2	3	4	5
Playing games online with others (Multiplayer Online Games)	1	2	3	4	5
If I play with others on the internet,					
<i>ØLeave this question open if you do not play with others on the internet!</i>					
...I use English in my In-Game chats	1	2	3	4	5
...I speak English through TeamSpeak, Mumble or comparable services	1	2	3	4	5
I use search engines such as Google in English	1	2	3	4	5
I use English apps on the phone	1	2	3	4	5

Therefore, the order of the items was altered in the final version of the EE questionnaire, looking like this:

(26)

English online					
I use English apps on the phone	1	2	3	4	5
I use search engines such as Google in English	1	2	3	4	5
Playing games on the computer or on gaming consoles	1	2	3	4	5
Playing games online with others (Multiplayer Online Games)	1	2	3	4	5
⇒ If I play with others on the internet,					
<i>➤Leave this question open if you do not play with others on the internet!</i>					
...I use English in my In-Game chats	1	2	3	4	5
...I speak English through TeamSpeak, Mumble or comparable services	1	2	3	4	5

This section was followed by the multi-item scale for reading in English, which included the following genres in the first draft of the questionnaire:

- books
- audiobooks
- articles in newspapers, magazines or news portals
- information texts, such as instructions or recipes
- short stories
- fanfiction
- comics or manga
- blogs or forum entries
- e-mails
- social media messages
- status messages or comments
- SMS or WhatsApp messages

Since the reliability for this scale was satisfactory with 0.84 it was decided to keep all of the items for the final EE questionnaire, which was used for the 2nd and 3rd data collection (see Table 6.2).

A similar procedure happened for all items concerned with the sub-construct of engagement with English writing, with the following genres being included in the first and final version of the EE questionnaire:

- stories
- fanfiction
- comics or manga
- blogs or forum entries
- e-mails
- social media messages
- status messages or comments
- SMS or WhatsApp messages
- lists or notes
- diary

The multi-item scale related to English writing achieved a reliability score of 0.86, which was considered satisfactory for a 10-item construct.

Concerning EE activities related to speaking English different modalities were considered, such as speaking English via Skype or other internet services, via phone or in face to face interactions. Additionally, talking or thinking to oneself in English was also considered, as well as using English words in other languages. Considering the piloting results the items were reduced and rephrased as follows:

(27)

Speak English					
I speak English on Skype or other internet services	1	2	3	4	5
I speak English on the phone	1	2	3	4	5
I speak English in face to face interactions	1	2	3	4	5
I think or talk to myself in English	1	2	3	4	5

While the item related to the integration of English into other languages had to be deleted to increase the reliability value of the scale from 0.74 to 0.81, one item had to be rephrased. Item three *I speak English with other people* was reformulated as *I speak English in face to face interactions* to clarify and sharpen its meaning in the final EE questionnaire.

Finally, the last defined sub-construct concerned with active and passive engagement with English theatre plays was deleted in the final EE questionnaire due to the lack of interest in these activities. However, to account for niche activities, the following open question was included in the final draft of the EE questionnaire:

(28)

I do other activities in English, namely: ➡ _____	1	2	3	4	5

Moving on to the third construct of the questionnaire, interested in the recording techniques used to collect vocabulary in the spare time, revealed that the construct, which consisted of only three items in the first draft, resulted in a relatively low reliability score of 0.49 the piloting. Consequently, it was decided that the construct either has to be excluded completely or has to be extended to increase its value. Therefore, in the final version of the EE questionnaire the original three items, *I collect new words in any written form*, *I remember new words without writing the down* and *I do nothing* were extended by two more items resulting in the following scale:

(29)

	(nearly) never				(nearly) always
I do not write down new words in an organized form. I just repeat the words orally.	1	2	3	4	5
I do write down new words but not in any organized form.	1	2	3	4	5
I keep a word list with new words	1	2	3	4	5
I make flash cards with new words	1	2	3	4	5
I keep an additional private vocabulary folder	1	2	3	4	5
I do nothing	1	2	3	4	5

Finally, the fourth construct was composed of 21 items related to Nation's (2013) taxonomy of word knowledge¹⁶. The following Table 6.3 tries to illustrate how the items are supposed to match to parts of the taxonomy (Nation, 2013):

I consider the spelling of a new word.	➡	Written word form
I repeat a new word orally and practice pronunciation.	➡	Spoken word form
I look up the phonetic transcription of a new word.	➡	Spoken word form
I try to divide the word into word parts that I understand.	➡	Word parts
I try to find words in other languages that help me.	➡	Word parts
I look up the members of the word family.	➡	Word parts
I search for translations.	➡	Form-meaning link
I look up definitions.	➡	Form-meaning link
I look up in which different contexts the word is used.	➡	Concepts and referents
I look for different shades of meaning of a word.	➡	Concepts and referents
I search for synonyms.	➡	Associations
I search for antonyms.	➡	Associations
I use the word in a sample sentence.	➡	Grammatical functions
I search for the word class of the word.	➡	Grammatical functions
I look up the correct grammatical usage of the word.	➡	Grammatical functions
I look for frequent collocations of the word.	➡	Collocations
I look up the level of formality of a word.	➡	Constraints of use
I look up the currency of the word in modern language use.	➡	Constraints of use
I look up whether the word has meaning differences in varieties of English, such as British English or American English.	➡	Constraints of use
I think about the level of abstraction of the word.	➡	Constraints of use
I look up whether the word is restricted to a specific domain, such as law or medicine.	➡	Constraints of use

Table 6.3: Link EE items with Nation's (2013) taxonomy

It has to be emphasized at this point that the connection of the actions taken by the language users and the word knowledge aspects they might be related to are made purely on basis of theoretical reasoning. Nation's (2013) taxonomy is working with a distinction of receptive and productive knowledge of a word knowledge aspects. However, there is no

¹⁶ See chapter 2.4.2.1 for more details on Nation's (2013) taxonomy.

such differentiation in the items of the EE questionnaire, since while Nation (2013) is talking about receptive and productive KNOWLEDGE, the items do not tap into measuring knowledge but are focusing on actions taken by the participants towards increasing one of the word knowledge aspects. However, these actions do allow to make inferences whether receptive or productive word knowledge is trained. Consequently, it might be interesting to see empirically into which aspects of word knowledge certain actions of learners might tap into.

As visible in Table 6.2 the scale resulted in a satisfactory reliability score of 0.84. Consequently, it was decided to keep all items due to their compositional and central nature.

As the last point in the questionnaire factual questions (Dörnyei, 2007, p. 102) on students biographical background encompassing specific open questions on participants' language background, age of onset, experience abroad, gender and age were included. No changes were made in this section from the piloting to the final draft of the EE questionnaire.

Finally, a post hoc item analysis was conducted after the administration of the EE questionnaire in the second data collection in June. For this purpose exploratory factor analyses were used. However, these results can be found in chapter 9 together with the results related to EE behaviour.

Referring back to Figure 6.2 it can be seen that 110 EE questionnaires were completed at each of the two data points when it was utilized. This amounts to the data of 220 EE questionnaires, which was entered and coded in the program SPSS (IBM, 2020). Similar as for the data related to the tests on academic verb knowledge the data entry was double checked and monitored by a second colleague.

6.3 Participants

Considering the longitudinal nature of this research study described above and the amount of data collected within one academic year, it is apparent that there was a considerable dropout rate of participants over time.

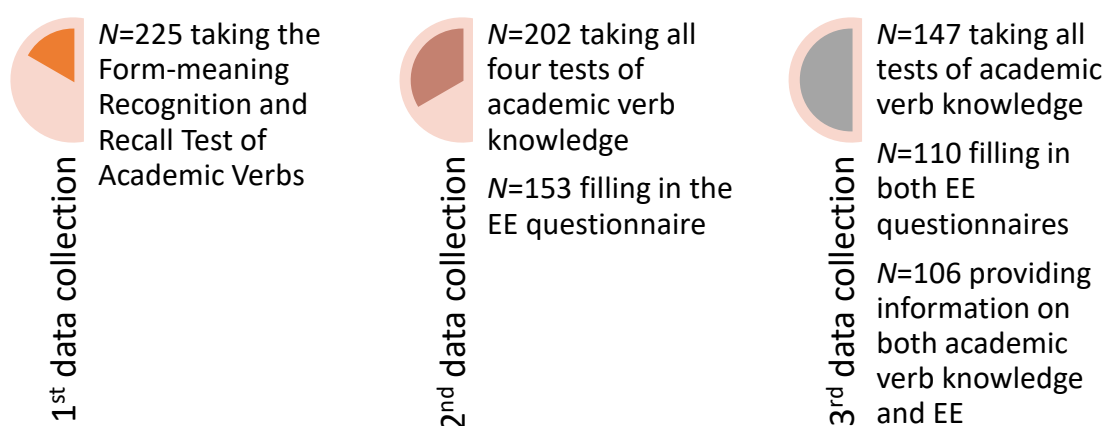


Figure 6.4: Visualization of drop-out rate

As visible in Figure 6.4 above the study started with 225 English major students of the University of Vienna, who were in the second semester of their studies. All students had just passed their entrance exams at the end of the first term and enrolled in their first seminar focusing on language competence. Moreover, participants had completed the Vienna English Language Test (VELT), which ensures a high B2 English language level of all students (VELT, 2022). Participants were recruited during regular class time and also the participation in the study was mostly taking place within official class time. The only exclusion criteria was that German had to be at least an educational language. However, since all students need to fulfil a minimum level of A2 according to the Common European Framework of Reference (CEFR) to apply for admission to a Bachelor program at university of Vienna this precondition did not prevent many students to participate in the study.

Out of the students participating three quarters of the group were studying in the teachers degree program to become English teachers for secondary schools and about one quarter of the group were enrolled in the BA English studies focusing on Linguistics.

All students were informed about the research study via a form of consent (see appendix 6.5) and by the researcher in person. Only students who agreed to participate via the form of consent were included in the study. To ensure anonymity of participants within the study, while being able to link all tests and EE questionnaires collected to one person to allow making statements about longitudinal developments and links between the two concepts researched an anonymous code was used. Both the EE questionnaire and all academic verb

tests contained a set of instructions to construct a five-digit code following the example of Kearney et al. (1984) as cited in Dörnyei (2010: 81–82). Example 30 shows the instructions as they were given on all the data collection instruments used.

(30)

My Code:			1st und 2nd letter of your first name
			1st letter of your mothers first name
			month of your birth (example: 03 for March)
Example: MAS06 (Maria; Susi; June)			

Moreover, participants were informed that they could withdraw from the project whenever they wanted, which is one of the potential reasons for the reduction of participant numbers over time. After one semester only 202 of the original 225 students filled in all four tests of academic verb knowledge and only 153 out of these 202 handed in the filled in EE questionnaire. While the drop out regarding the tests was still relatively low after the first semester only 147 out of the 202 students from the second data collection completed the vocabulary tests at the third data collection. Moreover, only 110 handed in the completed EE questionnaire and 106 provided the complete data set.

Reasons for the reduction of participant numbers can be partly attributed to the overall drop-out rate of participants ending their studies during the time of investigation. Since the group observed was still rather at the beginning of their studies some students still decided that they did not want to continue studying English. Moreover, some students failed the language competence course in the first semester, preventing them from participating in the next course in the following semester. Therefore, these participants also had to be excluded. Furthermore, some participants did not hand in parts of the data or were absent at the time the data collection took place. Although extra sessions were offered for people who were ill, this offer was only accepted very reluctantly. Finally, some participants forgot their codes used for anonymisation of the data and making it impossible to relate their subsets to each other.

Moving away from participants numbers to the biographical features of the student group observed, one can see in Table 6.4 that the gender of participants is rather unevenly distributed. This gender imbalance is, however, characteristic for the gender distribution in the teachers program for English as well as for the Bachelor of English Studies.

English major students (N = 153)			
Gender	female	N = 117	77%
	male	N = 35	23%
Mean age	in years	M = 21.24 [20.65, 21.82]	SD = 3.66
Level of English	acc. to CEFR	B2-C1	
Age of onset	in years	M = 7.46 [7.09, 7.84]	SD = 2.36
Experience abroad	Yes	N=37	24.3%
	No	N=115	75.7%
Number of languages spoken	2	N = 67	43.8%
	3	N = 50	32.7%
	4	N = 26	17.0%
	5+	N = 10	6.5%
Highest educational degree	School leaving exam	N=144	94.1%
	Bachelor degree	N =5	3.3%
	Master degree	N = 3	2.0%
	Doctor degree	N = 1	0.7%

Table 6.4: Biographical profile of participant group

Regarding the mean age of the participant group students were on average 21.24 years old when participating in the study and had proof of B2-C1 English language level according to the CEFR. While most students started their official English training around the age of 7.46, which corresponds to the age when English is introduced in Austrian primary schools, some students (N=16) indicated that they started learning English between the age of 0 and 4, which might be interpreted as being raised bilingually. Moreover, a relative large number of 60 students reported that their English instruction had started even later between the age of 9-12. This age range would correspond to the transition period from primary to secondary school in Austria. English was officially integrated as *verbindliche Übung* in the Austrian primary school curriculum in past years, meaning that students did not receive grades in this subject but should receive integrated English instruction of at least 32 hours in the first two grades and one hour per week in third and fourth grade (Lehrplan der Volksschule, 2005). However, intensity of language instruction varied considerably from

school to school at primary school level, which might explain the perceived late start of English instruction for this participant group.

Taking a look at participants highest educational degree at the time of participation within the study the clear majority (94.1%) reported to have a school leaving exam either from a secondary school (72.5%) or a vocational college (21.6%). While secondary schools in Austria have the main aim of conveying a thorough general education in eight years, vocational colleges comprise five years of first vocational additional to general education starting from the ninth grade onwards. Only 5.9% of the participants had a tertiary degree. Consequently, the group appears to be rather homogeneous regarding their highest educational degree achieved.

In comparison to their educational experience with English the participant group appeared to be rather homogeneous regarding experience of living abroad in an English speaking country with 75.7% of the students having no international experience. While experiences abroad are nowadays an integrative part of studying, the absence of stays abroad does not come as a surprise considering that the participants are rather at the beginning of their studies. Consequently, the 24.3% of the participants who already had experience abroad gathered it during their school or transition time between secondary and tertiary education.

Unlike experience abroad, participants displayed a rather large diversity of the number of languages spoken. While the majority of the group (43.8%) reported to speak only two languages, 32.7% were speaking three, 17.0% percent four and 6.5% five or more languages. Overall, 28 languages were reported to be spoken by the participant group.

7 Test development

The following chapter gives details on the development of the four vocabulary tests on academic verb knowledge used in this study. This comprises the Form-meaning Recall Test of Academic Verbs, the Form-meaning Recognition Test of Academic Verbs, two formats of the Collocation Recognition Test of Academic Verbs and the Multiple Meaning Recognition Test of Academic Verbs. Consequently, students were tested on their form-meaning recognition and recall knowledge of academic words as well as collocation recognition and multiple meaning recognition knowledge. All tests and their development are described in more detail below. In addition, taking up Gyllstad's (2019) suggestion a test manual¹⁷ containing information on the target group, purposes, context reliability and validity evidence and scoring procedure is provided together with the tests and their solutions in appendix 7.1.

¹⁷ The test manual of the Form-meaning Recognition Test of Academic Verbs can be found in appendix 7.1.5. Of the Form-meaning Recall Test of Academic Verbs the test manual can be found in appendix 7.1.10. The test manual of the Collocation Recognition Test of Academic Verbs can be found in appendix 7.1.15 and the test manual of the Multiple meaning Recognition Test of Academic Verbs can be found in appendix 7.1.19.

7.1 Form-meaning Recall Test of Academic Verbs

This chapter will describe the rationale and procedure of developing and improving the Form-meaning Recall Test of Academic Verbs. This comprises preliminary theoretical considerations about test design, reasons behind item selection as well as information about piloting and conceding adaptations made.

7.1.1 Basic considerations

To specify the aims of the test Read's (2000, pp. 7-13) "three dimensions of vocabulary assessment" are used as an aid to focus test design and are illustrated below in Table 7.1.

Discrete tests are said to define vocabulary knowledge as a measurable dimension separated from other language components, while embedded vocabulary tests link their testing construct to the relationship of vocabulary knowledge to abilities in other areas of language competence (Read, 2000, p. 8). Hence, embedded tests have a broader construct, while discrete tests strive to avoid bias through other areas of language ability.

The second differentiation is related to item selection of a test. In selective tests the test designer has chosen a particular group of words to be included in the tests for specific reasons (Read, 2000, p. 10). On the other hand, comprehensive tests strive to measure the whole scope of a test-taker's vocabulary, for example through the analysis of the words used by a participant in a written or spoken utterance.

Finally, Read (2000, p. 11) addresses another important issue in test design, namely the difference between context independent and context dependent tests. While context independent tests measure if a test-taker can produce or recognize a word without surrounding information, context dependent tests require the participant to make use of the context to elicit the word (Read, 2000, p. 11). A wide debate has been going on whether it is better to test participants' knowledge of words in or out of context.

While early vocabulary tests wanted to ensure construct validity by presenting the target words without context, more recent studies have raised awareness of certain difficulties of testing vocabulary in isolation. First of all, polysemy cannot be accounted for, if words are presented without surrounding context (Read, 2000, p. 99). Moreover, second language researchers are often interested in students' ability to use surrounding information for inferencing (Read, 2000, p. 100). However, few studies investigating the role of context have been conducted, which means that so far, no testing method can be preferred (Read, 2000, p. 100-101).

Discrete A measure of vocabulary knowledge or use as an independent construct.	vs.	Embedded A measure of vocabulary which forms part of the assessment of some other larger construct.
Selective A measure in which specific vocabulary items are the focus of the assessment.	vs.	Comprehensive A measure which takes account of the whole vocabulary content of the input material [...] or the test-taker's response [...].
Context independent A vocabulary measure in which the test-taker can produce the expected response without referring to any context.	vs.	Context dependent A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response.

Table 7.1: Dimensions of Vocabulary assessment (Read, 2000, p. 9)

Keeping these dimensions in mind we can sharpen the aims of the Form-meaning Recall Test of Academic Verbs. Since one of the goals of this study is to measure whether there is a relationship between Extramural English as an approach to vocabulary learning and academic vocabulary development, it appears to be sensible to choose a discrete construct, which pictures vocabulary knowledge as a separate category. Nevertheless, this does not mean that the importance of investigation of the fuzzy boundaries between different language components, such as grammar and vocabulary, are negated. However, this would exceed the scope and focus of this study.

Foreshadowing item selection, which will be discussed in more detail later, the Form-meaning Recall Test of Academic Verbs is a selective test, since the words tested have been chosen with particular reasons by the test designer and do not depict the whole scope of vocabulary used by a test-taker. One of the main reasons for this is the specific focus of this study on academic verbs, for which eliciting methods, such as the analysis of a test taker's response seem to be less suitable. Moreover, the test needs to be quick to administer, due to time constraints during data collection.

Regarding test design the Form-meaning Recall Test of Academic Verbs is difficult to place in any of the two categories. Factually the words are presented in minimal context and are not isolated, which classifies the test as context dependent. However, one needs to be aware that the context is not very rich and has been simplified as outlined in more detail in later sections.

In sum, the Form-meaning Recall Test of Academic Verbs is a discrete, selective and context dependent test.

7.1.2 Item selection

Item selection was restricted to academic words. All participants of the study are in their initial phase of their English studies and need to focus on and adapt to academic language use. Therefore, it appears to be a valid argument that overall vocabulary size is not the main and only interest for these language learners, but that vocabulary is learned with the particular purpose of academic language use in mind, requiring a specific vocabulary learning focus.

As Read (2000, p. 87) and Nation and Webb (2011) mention, using a word list for sampling is one of the most straightforward strategies. Hence, all words included were taken from the Academic Word List (AWL) developed by Coxhead (2000), which was already described in more detail in chapter 2.3.5. Although the AWL is not the most recent and advanced list (Gardner & Davies, 2014, p. 307) it serves as a reasonable source of words for this study, since students are introduced to the AWL in their courses. Therefore, the AWL is the most familiar academic word list for the students.

Initially the test integrated nouns and verbs from the AWL, since reporting verbs (deChazal, 2014; Luzon, 2018, p. 175) and nominalization are crucial features of academic texts, which require a broad knowledge of academic verbs and nouns (Charles, 2014; deChazal, 2014; Graham, 2018, p. 7). However, nouns were excluded after further consideration due to the diverse nature of words which co-occur with academic nouns. Hence, it would have been difficult to create suitable and regular items for the collocation recognition test.

In a second step cognates were excluded from the test items, since they had nearly identical equivalents in German and would have been too easy for the target group. Nation (1990, p. 262) addresses this issue concerning his Vocabulary Levels Test, which includes a section on the University Word List. He argues that languages with a close relationship to Latin have too many similarities to the target words and therefore, these cognates would influence the scores observed.

Coxhead (2000) ranks all words on the AWL according to frequency of occurrence in academic texts resulting in ten frequency bands, all covering 60 words except the lowest frequency band, which covers only 30 items. To ensure near equal distribution amongst frequency levels, ten academic verbs were chosen for eight frequency bands of the AWL. It is only for two frequency bands (Sublist 6 and Sublist 8) that eleven and nine words were selected, since there were not enough eligible verbs in Sublist 8. For the last frequency band, five words were chosen, since it covers only half of the number of words of other frequency bands. All verbs were selected with the help of a random number generator (Haahr, 2020). This adds up to a total number of 95 words, which were used for the test. Table 7.2 lists all selected verbs.

Sublist 1 (10 words)	assume, approach, assess, estimate, distribute, consist, establish, derive, require, benefit
Sublist 2 (10 words)	restrict, maintain, affect, purchase, perceive, reside, focus, achieve, acquire, injure
Sublist 3 (10 words)	contribute, constrain, justify, ensure, shift, rely, remove, deduce, fund, consent
Sublist 4 (10 words)	predict, approximate, commit, attribute, impose, label, occupy, retain, emerge, promote
Sublist 5 (10 words)	amend, evolve, decline, facilitate, pursue, reject, sustain, alter, aware, expose
Sublist 6 (11 words)	utilize, acknowledge, inhibit, enhance, reveal, presume, exceed, allocate, assign, attach, display
Sublist 7 (10 words)	submit, insert, dispose, comprise, prohibit, release, deny, infer, transmit, confirm
Sublist 8 (9 words)	displace, append, deviate, contradict, abandon, terminate, restore, induce, reinforce
Sublist 9 (10 words)	diminish, confine, found, devote, assure, trigger, coincide, restrain, cease, attain
Sublist 10 (5 words)	conceive, undergo, invoke, encounter, assemble

Table 7.2: Item selection 1st draft

7.1.3 Item development

Test items were developed for all of the 95 academic verbs chosen. Similar to the Productive Vocabulary Levels Test (Laufer & Nation, 1999) each word is elicited with the help of a sentence. However, the context in which the target words are presented are as “neutral” as possible. This means that the surrounding context should not give away possible answers but the presentation in minimal context restricts which meaning aspect of a word is required. As example 24 shows, presenting the word *assume* in the context of a sentence clarifies whether the underlying meaning *to suppose* or *to undertake* is requested. An additional reason for keeping the context rather minimal is that reading skills of students should not interfere with their scores.

(24) *They might as_____ that it came from taxpayer dollars.*
(*annehmen*)

To ensure that the likelihood of giving the correct answer is not influenced by the difficulty of the words in the sentences, a frequency analysis¹⁸ of the test was conducted with the updated version of the program LexProfiler (Cobb, 2021) called Compleat Web VP, which is

¹⁸ The lexical profile of the Form-meaning Recall Test of Academic Verbs can be found in appendix 7.2.1.

based on the New General Service List (Browne, 2013). The lexical profile of the test clearly indicates that the context provided should be easily understandable for English major students, since 82.1% of all words were amongst the most frequent 1000 words of the English language. 9.9% of the words were amongst the most frequent 2000 words and 5.0% belonged to the K-3 level. Hence, over 95% of all words in the test's context can be found in the most frequent 3000 words of the English language. This equates the threshold of token coverage necessary to understand a text mentioned by Nation (2013, p. 206). Only, 3% of all the words in the context can be attributed to a higher frequency band. Table 7.3 shows the cumulative token percentage for the individual frequency bands with the threshold of over 95% highlighted.

Frequency level	Cumulative token percentage
K-1	82.1%
K-2	92.0%
K-3	97.0%
K-4	98.1%
K-6	98.4%
K-9	98.6%
Off list	100%

Table 7.3: Lexical profile of Recall Test of Academic Verbs

Beneath the sentence the German equivalent for the word to be elicited is given (see example 24). This bilingual approach is necessary for two reasons. Firstly, students' form-meaning recognition AND recall knowledge of the same items is tested in this study. Students cannot be given the recognition test before the recall test, since the synonyms and definitions provided as options in the recognition test might already help them to answer the recall items. Hence, the Form-meaning Recall Test of Academic Verbs is administered before the Form-meaning Recognition Test of Academic Verbs. However, giving a synonym beneath the sentence of the Form-meaning Recall Test of Academic Verbs would already give away a possible answer of the Form-meaning Recognition Test of Academic Verbs. Therefore, the only possible solution is using a German translation. Of course this means that the group of participants needs to be even more homogenous, since only students with German as main educational language can be included. However, this has already been addressed in chapter 6 describing participants' profiles. The second argument for a recall test incorporating two languages is that the items should have the same prerequisites for being demanding. To explain this further two sample sentences taken from Read (2000, p. 125) will be discussed:

(25)a) The picture looks nice; the colours blend really well. [own emphasise]

b) Many companies were manufacturing computers. [own emphasise]

To test the underlined words several letters are deleted. This works fine with sentence a) where the word tested is unambiguous when the first two letters remain.

(26)a) The picture looks nice; the colours bl_____ really well.

Nevertheless, this does not work for sentence b), since the first two letters of the word tested do not limit possible answers effectively. There are more words, which would fit in the blank, such as *making* or *maintaining*.

(26) b) Many companies were ma_____ computers.

Therefore, more than two letters need to be given. For some words this could even mean that already half of the word needs to be given to restrict answers effectively. Thus, items with more letters given are automatically less demanding than words with only two predetermined letters. Read (2010, p. 125) states that this fact blurs the unambiguousness of what the test design is measuring. Providing a L1 equivalent eliminates this problem, since the semantic options are further restricted by the translation given. Consequently, the translation component does alter a crucial aspect of the design, namely the role of inferencing, as a side effect. While the original design by Laufer and Nation used in the active version of the Vocabulary Levels Test (Laufer & Nation, 1999; Laufer & Nation, 1995) incorporates inferencing in the test concept, the revised version with L1 translation eradicates the option of guessing from context. The context provided by the sentence becomes less relevant for word choice, since the meaning has already been predetermined by the L1 word given. Therefore, reading skills are less significant for test performance. This consequence might not be desirable for Laufer and Nation's studies (1999, 1995), but it suits the theoretical construct of this research. Since this study investigates the relationship between learning behavior and vocabulary knowledge, it is of utmost importance that the concept measured by the tests employed is restricted to vocabulary knowledge without bias of other skills, such as reading or writing.

Scoring

Concerning the scoring of the Form-meaning Recall Test of Academic Verbs it was decided that each correct answer will be awarded with one point, while missing and incorrect answers will be treated identically with 0 points. The reason for not differentiating between missing and incorrect answers is that for the research questions investigated only the actual accurate knowledge gain is interesting and not whether participants did not know the word or their answer was incorrect. In the Form-meaning Recall Test of Academic Verbs attention is paid not only to the appropriate word meaning choice but additionally to correct word form and grammatical use of the word in the given sentence. Hence, it cannot be

considered to be a pure recall test of the form-meaning link but it builds on productive knowledge of word form and grammatical use.

Moreover, it has to be highlighted that the Form-meaning Recall Test of Academic Verbs is not designed to give insights into pupils' overall productive vocabulary size, since it only covers a maximum of ten items per frequency band, which is too little to extrapolate reliably (Gyllstad, McLean & Stewart, 2021; Gyllstad, Vilkaite & Schmitt, 2015). As Gyllstad, McLean and Stewart (2021, p. 571) outline the number of items per frequency band included in a test considerably influences the exactness of the results achieved by extrapolation, as shown through the development of confidence intervals.

7.1.4 Piloting¹⁹

Following the recommended steps for conducting a well-founded pilot study, (Dörnyei, 2007, p. 112-113) all items were reviewed informally with a group of colleagues and instructions, explanations and examples were reviewed by a native speaker of English. After these initial refinements the Form-meaning Recall Test of Academic Verbs was first piloted in January as can be seen in chapter 6.2.

Research questions

The aim of the piloting was to answer the following questions:

1. Is the Form-meaning Recall Test of Academic Verbs satisfactory concerning difficulty?
2. Is the Form-meaning Recall Test of Academic Verbs reliable?
3. What can be stated about response validity of the Form-meaning Recall Test of Academic Verbs?

These questions need to be answered in order to refine the test and additionally to help reduce items.

Participants

The informants of the piloting were 36 English major students, who were recruited from an optional tutorial for a lecture called *Grammar in Use*. All participants were comparable to the target group of the main study and were not specifically prepared for the tests. The group consisted of three male and 33 female students, who were mostly aged between 18 and 20.

Test administration

Participants were asked to fill in the Form-meaning Recall Test of Academic Verbs and then proceed to the Form-meaning Recognition Test of Academic Verbs, which will be described

¹⁹ The piloting version of the Form-meaning Recall Test of Academic Verbs can be found in appendix 7.1.6 together with its solution.

in more detail later. Students were advised not to spend too much time thinking about individual items. Instead they should skip an item, if they cannot remember the answer quickly, as an attempt to minimize guessing effects. Participants were informed that guessing would be treated more strictly than not knowing. On average students took 60 minutes to complete both tests. Students also filled in an extra questionnaire on the tests that gave them the opportunity to provide me with anonymous feedback on the tests.

Results

The results of the piloting were generated in two steps. First descriptive statistics for the Form-meaning Recall Test of Academic Verbs were calculated, including the mean (M) and the standard deviation (SD). In a next step an estimate of test score reliability (Cronbach's alpha) as well as item facility values and discrimination values were considered (Green, 2013, p. 28). Lastly, information concerning face validity of the test format, such as results of the questionnaire were given.

As can be seen in Table 7.4 participants could achieve a maximum score (s) of 95 points. Given that the data is normally distributed, $D(36)=0.951$, $p=0.115$, it is most informative to look at the mean score (Dörnyei, 2007, p. 214). In this case the mean was 37.53 with a standard deviation of 15.90, which is below one third of the total score. Hence, the test seems to be relatively difficult for students. Since we have a large standard deviation the group of participants appears to be rather heterogeneous, which means that some students are more successful than others in completing the test correctly.

Value (s)	Total (95)
M	37.53
SD	15.90
α	0.93

Table 7.4: Descriptive results on Form-meaning Recall Test of Academic Verbs (N=36)

The reliability of the Form-meaning Recall Test of Academic Verbs was relatively high with a Cronbach α of 0.93 (Field, 2018, p. 823). Hence, the test appears to be reliable (Field, 2018, p. 821). However, still improvements can be made as is reflected in the calculations how the score would vary, if particular items were deleted. However, since Cronbach α is influenced by item number, with a fewer items resulting in a lower score (Field, 2018), also item facility and discrimination values have been taken into account.²⁰ Based on these statistics, 6 items namely *deduce* (item 9), *injure* (item 14), *consent* (item 30), *expose* (item 47), *transmit* (item 62), *confirm* (item 68) were deleted. Moreover, item 80 (*exhibit*) had to be deleted due to

²⁰ For the SPSS results regarding Cronbach alpha, as well as exact numbers for facility and discrimination values see appendix 7.3.1.

ambiguous answer options in both tests and *item 4 (benefit)*, *item 12 (achieve)* and *item 26 (remove)* were deleted, since they had zero variance, meaning that all participants knew the answers to these items already. Additionally, five more items, namely *require (item 8)*, *acquire (item 17)*, *fund (item 24)*, *attach (item 57)* and *aware of (item 42)* were deleted, since their exclusion did not reduce the reliability score of either test and the test was still too long covering 85 items. In summation, after thorough consideration and comparison to the α -scores of the Form-meaning Recognition Test of Academic Verbs, 15 items were deleted, namely *benefit*, *require*, *deduce*, *achieve*, *injure*, *acquire*, *fund*, *remove*, *consent*, *aware of*, *expose*, *attach*, *transmit*, *confirm* and *exhibit*.²¹ The exclusion of these items resulted in a Cronbach α of 0.94.

Regarding face validity participants reported no difficulties with particular items²². However, it was mentioned repeatedly that the tests were too long and that students had to fight against test exhaustion and fatigue, which corresponds well to the need to exclude some of the original test items as mentioned above. Concerning perceived difficulty of the test students were asked to rate the degree of difficulty of each test on a seven-point Likert scale. Since the distribution was not normal [$D(34)=0.886$, $p=0.002$], the median and the interquartile range should be interpreted (Bachman, 2004, p. 77; Dörnyei, 2007, p. 214). The median score for the Form-meaning Recall Test of Academic Verbs was 6, with an interquartile range of 1, which indicates that students conceived the test to be doable but relatively difficult and that their evaluation of the test does not seem to vary strongly, which is also confirmed by the extreme scores reported, which reached a maximum of 7 and a minimum of 4.

N	34
M	5.5
Mdn	6
SD	1.05
Q	1.0

Table 7.5: Descriptive results for Form-meaning Recall Test of Academic Verbs questionnaire (N=34)

²¹ Since both the Form-meaning Recall Test of Academic Verbs and the Form-meaning Recognition Test of Academic Verbs comprise the same items, due to their purpose of measuring different components of word knowledge for the same academic verbs, reduction of items was a negotiation between both tests. This means that the decision which items were to be deleted had to be taken in consideration of the reliability value of both tests. Consequently, item 78 (abandon) was left in the Form-meaning Recall Test of Academic Verbs although it could have been deleted, since it would have lowered the reliability coefficient of the Form-meaning Recognition Test of Academic Verbs more than it would have helped to improve the Form-meaning Recall Test. Therefore, it might be possible that reliability values of one test decrease slightly in an effort to improve the second one.

²² For more detailed insights into the SPSS output regarding the face validity questionnaire see appendix 7.3.2.

Comparing the median of students' test-evaluations ($Mdn=6$) with the mean score reached in the piloting ($M=37.53$) indicates that students expect their performance on the test to be relatively similar to what it actually was. This factor implies that face validity of the test is satisfactory. However, one possible option to increase face validity even more might be to include more elaborate instructions to draw attention to the necessity of correct spelling and grammar.

7.1.5 1st data collection²³

As described in chapter 6.2 the refined Form-meaning Recall Test of Academic Verbs, as described above, was employed in March in the first data collection.

Research questions

Apart from being the first data collection point of this study aiming at answering the research questions of this PhD, data was used to answer the following questions about the test:

1. Is the difficulty of items adequate for the group of test takers?
2. Is reliability of the Form-meaning Recall Test of Academic Verbs satisfactory?
3. Do the items of the Form-meaning Recall Test of Academic Verbs discriminate satisfactorily between successful and less successful participants?
4. Should anymore items be reduced?

Test administration

Participants were given the Form-meaning Recall Test of Academic Verbs first. A short oral instruction was given as well as the possibility to ask open questions. All students were informed that they should not spend too much time on one item and instead of lingering they should skip the item and move on to the next one. After all subjects had filled in the Form-meaning Recall Test of Academic Verbs these test booklets were collected and the Form-meaning Recognition Test of Academic Verbs was distributed.

The maximum time available to fill in the tests was 45 minutes in total. This amount of time was shaped by students' statements in the piloting and the institutional embedding of the study during regular class time. However, overall 45 minutes even seemed to be too much time to complete the tests, since students were finished earlier and reported that they could have completed the tests in a shorter time. Nevertheless, students varied strongly in how they distributed the time between the two tests. While some of them spent 35 minutes on the Form-meaning Recall Test of Academic Verbs and only ten minutes on the Form-meaning Recognition Test of Academic Verbs, other students spent an equal amount of time on the two tests. This variance was prevented in the next data collections, since the time

²³ The Form-meaning Recall Test of Academic Verbs used for the 1st data collection can be found in appendix 7.1.7 together with its solution.

spent on each test might function as a confounding variable. However, for the purpose of establishing a baseline of students' knowledge of academic verbs the administration of this data collection was sufficient.

Results

As in the piloting, results are generated in two steps, comprising descriptive statistics²⁴, such as calculating the mean and standard deviation of the test on the one hand and on the other hand a reliability analysis of the test under observation is conducted²⁵.

A Shapiro-Wilk test (Field, 2018, p. 251) revealed that the data was normally distributed, $D(225)=0.990$, $p=0.142$. As summarized in Table 7.6, students were able to reach a mean of 39.35 points of a total score of 80 points. Hence, the test seems to have been rather difficult on average for the participant group. The standard deviation, $SD=10.41$, indicates that the group was rather heterogeneous.

Value (s)	Total (80)
<i>M</i>	39.84
<i>SD</i>	10.56
α	0.88

Table 7.6: Descriptive results for Form-meaning Recall Test of Academic Verbs 1st data collection (N=225)

Comparing these descriptive findings to the numbers of the piloting [$M=37.53$, $SD=15.90$] reveals certain parallels. While the mean score is slightly higher in the 1st data collection [$M_{1st}=39.84$; $M_{pilot}=37.53$] the standard deviation has decreased insignificantly [$SD_{1st}=10.56$; $SD_{pilot}=15.90$]. Therefore, the target participant group appears to be slightly more proficient and homogeneous in their performance.

Item facility ranges from 0-1 and depicts the proportion of participants that answered an item correctly (Bachman, 2004, p. 126). An analysis of the facility values of individual test items, summarized in Table 7.7, revealed that 40% of all items were in the optimal range of 0.3-0.7 reported by Bachman (2004, p. 138), while 31% reached a facility value below 0.3, which means that they were rather difficult for the participants. While such a high level of difficult items might not be desirable in a standard test taken only once, it is beneficial for the longitudinal design of this study, since participants who will take the test three times

²⁴ Note that descriptive statistics calculated in this chapter for the sake of test development are not identical with the descriptive statistics (mean, SD...) found in chapter 8, describing the performance of the 147 participants, who participated in the study over the whole time of investigation. This is due to the fact that for the sake of test development results of all 225 participants, who were part of the study at this point were included, while the results described in chapter 8 are already considering the drop-out rate, which leaves 147 participants.

²⁵ For detailed results and SPSS output relating to the results presented in this chapter see appendix 7.3.3.

still have enough room for improvement (Popham, 2000). Moreover, the high number of difficult items makes it more likely that the test is still able to discriminate between successful and less successful students after the timespan of one year. Only 28% of all 80 items reached a score of above 0.7. These easy items have been retained for psychological reasons (Henning, 1987, p. 50). It has been shown that including easy items at the beginning of a test can help test takers with test anxiety to overcome their fear and show their full ability (Henning, 1987, p. 50). Detailed facility values for the individual items can be found in appendix 7.

above 0.7	28%
0.3-0.7	40%
Below 0.3	31%

Table 7.7: Distribution of facility values of Form-meaning Recall Test of Academic Verbs (N=225)

A reliability analysis of the Form-meaning Recall Test of Academic Verbs resulted in a Cronbach α of 0.88, which is clearly below the Cronbach α of 0.93 reached in the piloting with the same test (Field, 2018, p. 823). As all scores were double checked with a second scorer, mechanical errors in the scoring process can be eliminated as a possible explanation. Henning (1987, p. 77) additionally mentions differences in test administration as a possible reason for potential changes. Hence, the change in the administrative procedure regarding the maximum time to complete the two tests, can be seen as a potential influencing factor. As a reaction items were deleted by looking at the Cronbach α if item is deleted and discrimination values²⁶.

A common method to calculate item discrimination is Inter-item correlation, which basically already says what it is doing, namely it calculates the correlations between the scores on individual items and the total score of test takers (Henning 1987, p. 53). There are multiple ways to interpret discrimination values that can range between +1 and -1 but one of the most common approaches are the guidelines provided by Ebel (1979):

.40 and higher	very good items
.30 to .39	reasonably good items
.29 to .20	marginal items subject to improvement
below .19	poor items

Table 7.8: Categorization of discrimination values (Ebel, 1979)

²⁶ Item discrimination provides information about how well an item distinguishes between ‘good’ or ‘bad’ test takers related to the score (Green, 2013). In more detail item discrimination tells us whether an item is answered correctly mostly by test takers with a high overall score or not (Gyllstad, 2007). If an item is answered correctly by test takers with poor overall scores frequent then the item needs to be reconsidered.

A categorization of all items according to Ebel's (1979) guidelines (see Table 7.8) revealed that only 12.50% of all items had a very good discrimination score, 27.50% a good, 40% a marginal and 20% a poor discrimination score.

	Number of items	Percentage
<i>very good</i>	10	12.50%
<i>Good</i>	22	27.50%
<i>marginal</i>	32	40%
<i>Poor</i>	16	20%

Table 7.9: Discrimination values Form-meaning Recall Test of Academic Verbs (N=225)

While at first sight these results look not very reassuring, they can be explained and moderated. For instance, the discrimination score is influenced by the facility value of an item. If an item is answered correctly by 95% of all participants or is not solved by 95% of all participants, the item does not have much capacity to discriminate between “good” and “bad” participants. Hence, low discrimination scores can be partly explained by too easy or too difficult items of the test. Moreover, discrimination values are not to be treated as an absolute but as an inter-item comparative measure, since they are influenced by both sample size and ability range (Henning, 1987, p. 53). Nevertheless, the discrimination values should be increased overall and especially negative items need to be excluded.

On the basis of the discrimination scores and Cronbach α eight items, *restrict*, *focus*, *rely*, *decrease*, *abandon*, *reject*, *deny* and *restore*, were deleted. However, one needs to consider that if items are deleted from the Form-meaning Recall Test of Academic Verbs, they will be excluded in the Form-meaning Recognition Test of Academic Verbs as well and cannot be used in the Collocation Recognition Test of Academic Verbs or the Multiple Meaning Recognition Test of Academic Verbs anymore. Hence, as in the piloting the decision, item reduction is always a negotiation between the Form-meaning Recall and Recognition Test of Academic Verbs, as well as a matter of maintaining items with multiple meanings. Deleting the items mentioned above resulted in a stable final Cronbach α of 0.88 and the distribution of discrimination scores changed to 12.50% very good items, 29.17% good items, 40.28% marginal items and 18.06% poor items.

7.1.6 2nd and 3rd data collection²⁷

The second version of the Form-meaning Recall Test of Academic Verbs, now featuring 72 items, was utilized in the second and third data collection. No changes were made between data collection two and three. Additionally, the participants were identical to those in the first data collection.

Test administration

In the second data collection the Collocation Recognition Test of Academic Verbs and the Multiple Meaning Recognition Test of Academic Verbs were administered additionally to the Form-meaning Recall Test of Academic Verbs and the Form-meaning Recognition Test of Academic Verbs. All four test booklets were given to the participants at the same time. However, participants were instructed to start with the Form-meaning Recall Test of Academic Verbs, followed by the Collocation Recognition Test of Academic Verbs, the Multiple Meaning Recognition Test of Academic Verbs and lastly the Form-meaning Recognition Test of Academic Verbs. The order of the tests was chosen to avoid bias between the tests. Students should not receive hints on possible answers by the test they did before. Hence, the Form-meaning Recall Test of Academic Verbs was given first, since students have to provide the target words themselves and they were only given the German translations. The Collocation Recognition Test of Academic Verbs provided students with three possible word combinations, which pointed into different contextual directions. All three contextual directions, also the context of the distractors was resumed in the Multiple Meaning Recognition Test of Academic Verbs again. Therefore, students who did choose one of the distractors in the Collocation Recognition Test of Academic Verbs, would additionally find a distractor in the Multiple Meaning Recognition Test of Academic Verbs that points into the same (incorrect) direction. Lastly, students filled in the Form-meaning Recognition Test of Academic Verbs, since this was the test where most information about the target words was provided. The researcher and the lecturer of the group monitored whether the order was kept.

To avoid that students spend different amounts of time on the tests, a time limit was given for each of them. For the Form-meaning Recall Test of Academic Verbs students were given 15 minutes, for the Collocation Recognition Test of Academic Verbs eight minutes, for the Multiple Meaning Recognition Test of Academic Verbs 7 minutes and the Form-meaning Recognition Test of Academic Words had to be completed in 15 minutes again. Students were not allowed to move on to the next test before the timeslot was over and the previous test had been collected. Hence, all students spent the same amount of time on each of the tests and had a total amount of 45 minutes for all four tests. While this is the same amount of time for four tests as they had in the first data collection for only two tests, one needs to

²⁷ The Form-meaning Recall Tests of Academic Verbs used for the 2nd and 3rd data collections can be found in appendix 7.1.8 and 7.1.9 together with their solutions.

keep in mind that the Form-meaning Recall and Recognition Test of Academic Verbs were considerably shortened between the first and the second data collection and that students did not need the whole amount of time in the first data collection. Time reduction additionally also avoid threats to reliability and response validity through fatigue or boredom. After the administration of the second data collection students reported that they had no problems filling in all four tests in the given time.

Results²⁸

In the second data collection the Cronbach α increased to 0.90, which was most likely due to differences caused in the altered difficulty of items through learning that had taken place. A more detailed description of changes can be found in section 8.4. Another confounding factor of reliability could be the difference in administration from the first to the second data collection. However, influences like inter-rater variability or difference in test length can be ruled out. Changes in the true score of participants due to mere test effects or test-wiseness were prevented as far as possible by including a relatively large number of items in the test and by varying the order of the items from data collection to data collection.

7.2 Form-meaning Recognition Test of Academic Verbs

Similar to the Form-meaning Recall Test of Academic Verbs the Form-meaning Recognition Test of Academic Verbs is a discrete test focusing on vocabulary as an independent construct (Read, 2000). The same items as for the Form-meaning Recall Test of Academic Verbs were chosen, making it a selective test²⁹. Moreover, the Form-meaning Recognition Test of Academic Verbs is a context dependent test, since students are required to read and understand definitions and to match them to the correct target item.

7.2.1 Item development

Similar to the Form-meaning Recall Test of Academic Verbs the Form-meaning Recognition Test of Academic Verbs is based on the well-established design of the Vocabulary Levels Test devised by Nation (Nation 1983; Nation 2013). This format requires participants to match the target words with their correct definition. For each of the target words one suitable definition is given together with two distractors. Read (2000, p. 171) describes this as a recognition matching task “focusing on basic word meaning”, which classifies it as a suitable design for the purpose of this test. As common with this test format a number of target words are grouped together with their definitions and some distracting definitions (Read, 2000, p. 172). The distractor definitions were incorporated in the test to stop participants from getting correct answers through eliminating options. In example (27) four target words are given and the two distractors are highlighted in bold.

²⁸ More detailed SPSS output can be found in appendix 7.3.4.

²⁹ For more information on the exact item selection see section 7.1.2.

(27)

1. to carry something out	commit ____
2. to judge the quality or quantity of something	assess ____
3. to come near someone or something	approach ____
4. to contain something as a part of something else	establish ____
5. to add or mount up	
6. to set up on a strong and lasting basis	

Following Read's (2000, p. 172) advice the definitions have been inspired and selected from learner dictionaries, namely the Cambridge English Dictionary (University of Cambridge, n.d.), the Oxford English Dictionary (Oxford University Press, n.d.), the Oxford Learners Dictionaries (Oxford University Press, n.d.) and Linguee (DeepL GmbH, n.d.) to ensure understandability of the items. To decide which meaning aspect to include if multiple meanings were given was decided by comparing the meanings listed for each target word in these dictionaries. A meaning which was reported by all four dictionaries was included.

In the same context Read (2000, p. 172) emphasizes that the definitions used in the test "should be composed of higher frequency vocabulary than the words to be tested [...]", which means that for the purpose of this study the definitions should be composed of no academic words. To ensure this, all definitions were entered into the corpus analysis program LexTutor (Cobb, 2021), which generates lexical profiles for texts. This resulted in a coverage of 95.6% of all words being amongst the first two thousand words of the English language, ensuring a satisfactory coverage.³⁰

All words used in the definitions, which could not be found in the first 2000 words of the General Service List (West, 1953), were replaced by a highly frequent synonym. However, some exceptions were made, if the words were likely to be known by students, such as *conclusion*, no near synonym could be found, or if the words were followed by a near synonym or defining phrase in the definition. Table 7.10 below lists all words that were replaced.

Words outside 2K-level that were replaced	achieve, achieving, apparent, conclusion, display, enable, gradually, inadequate, initiate, intensify, interpret, involve, logical, obtain, occupied, perception, positively, remove, resources

Table 7.10: Words outside the 2000 frequency band Form-meaning Recognition Test of Academic Verbs

Grammatical or structural clues as mentioned by Read (2000, p. 120) were avoided by including only verbs. The order of items was based on frequency of occurrence with the

³⁰ For a more detailed lexical profile of the Form-meaning Recognition Test of Academic Verbs see appendix 7.2.2.

most frequent words appearing at the beginning of the test and the least frequent at the end. However, this order was changed during the data collection phases, similarly to the Form-meaning Recall Test of Academic Verbs, since adapted order of items might distract participants from the fact that they are tested on the same words repeatedly.

Concerning the scoring procedure of the Form-meaning Recognition Test of Academic Verbs students received one point if they managed to match the target word with the correct definition. Similar to the Form-meaning Recall Test of Academic Verbs students received 0 points, if the answer was incorrect or missing. Another accordance between the two tests is the restricted focus on academic verbs, which does not allow for extrapolation of vocabulary size estimates.

7.2.2 Piloting

The Form-meaning Recognition Test of Academic Verbs was reviewed by a native speaker of English, who paid attention to the clarity of instruction, examples and items. Moreover, the test was first informally taken by a small group of colleagues and friends, who volunteered to give informal feedback and who checked for typos and other formality issues. The Form-meaning Recognition Test of Academic Verbs was officially piloted in January as can be seen in chapter 6.2.³¹

Research questions

In the piloting the following questions were pursued:

1. Is the Form-meaning Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Is the Form-meaning Recognition Test of Academic Verbs reliable?
3. What can be stated about response validity of the test?

Participants

The participant group of the piloting of the Form-meaning Recognition Test of Academic Verbs was the same as for the piloting of the Form-meaning Recall Test of Academic Verbs.

Test administration

As described in section 7.1.4 the Form-meaning Recognition Test of Academic Verbs was administered after the Form-meaning Recall Test of Academic Verbs. A time limit of 60 minutes was given. After completing the Form-meaning Recall Test of Academic Verbs students were briefly introduced to the test's design and were given the opportunity to ask questions. As before students were asked to give feedback on the Form-meaning Recognition Test of Academic Verbs via a short questionnaire.

³¹ The piloting version of the Form-meaning Recognition Test of Academic Verbs can be found in appendix 7.1.1 together with its solution.

Results³²

As an analysis of the score distribution showed the data of Form-meaning Recognition Test of Academic Verbs was not normally distributed, $D(36)=0.940$, $p=0.052$. Therefore, the median and the mean are reported in Table 7.11 as measures of central tendency (Dörnyei, 2007, p. 214). As in the Form-meaning Recall Test of Academic Verbs the maximum score (s) was 95 points. On average participants reached a median score of 71.50 points and a mean score of 67.39 points. Hence, on average participants managed to reach a higher score on form-meaning recognition of academic verbs than on form-meaning recall. However, the semi-interquartile range of 30.5 indicates that the performance was rather varied.

Value (s)	Total (95)
<i>Mdn</i>	71.50
<i>M</i>	67.39
<i>Q</i>	30.5
α	0.96

Table 7.11: Descriptive results for Form-meaning Recognition Test of Academic Verbs (N=36)

The reliability coefficient of the Form-meaning Recognition Test of Academic Verbs was relatively high with $\alpha= 0.96$. However, similarly to the Form-meaning Recall Test of Academic Verbs one needs to consider that the reliability score is dependent on the number of items included in a test and tends to increase automatically when the number of items increases (Field, 2018, p. 821). Hence, the relatively high number of 95 items could be an influencing factor. At the same time reducing the total item number to 80 items by deleting the same 15 items as for the Form-meaning Recall Test of Academic Verbs³³ achieved a similarly high reliability coefficient of 0.96. Therefore, the Form-meaning Recognition Test of Academic Verbs reached a more than satisfactory reliability coefficient.

Similar to the Form-meaning Recall Test of Academic Verbs participants were asked to rate the difficulty of the test on a seven-point Likert scale and name any problematic items or inconveniences. Convergent to previous findings, participants complained that the test was too long and that they could not concentrate well enough. Therefore, the reduction of items is well justified by participants' feelings and by administrative constraints. Considering the difficulty of the Form-meaning Recognition Test of Academic Verbs participants rated the test to be of mid difficulty [$M=5.3$]. However, students' evaluations varied slightly with a standard deviation of 1.32. Normal distribution of the data was given, $D(34)=0.876$, $p=0.001$. The minimum rating of difficulty given for the Form-meaning Recognition Test of Academic

³² More detailed output from SPSS can be found in appendix 7.4.1.

³³ For more detail see chapter 7.1.4.

Verbs was 1, which was ticked by one person, while the maximum score was 7, which was chosen by at least five participants. All descriptive trends observed are summarized in Table 7.12.

N	34
M	5.3
SD	1.32

Table 7.12: Descriptive results for Form-meaning Recognition Test of Academic Verbs questionnaire (N=34)

Comparing these results to the piloting of the Form-meaning Recall Test of Academic Verbs, indicates that students perceived the Form-meaning Recognition Test of Academic Verbs to be slightly less difficult, which corresponds well with the average score reached.³⁴

7.2.3 1st data collection³⁵

Parallel to the administration of Form-meaning Recall Test of Academic Verbs the first data collection of the refined version of the Form-meaning Recognition Test of Academic Verbs took place in March³⁶.

Research questions

The data gathered has been used for further analysis and refinement of Form-meaning Recognition Test of Academic Verbs. The following questions were pursued:

1. Is the difficulty of Form-meaning Recognition Test of Academic Verbs items adequate for the group of test takers?
2. Is the reliability of the Form-meaning Recognition Test of Academic Verbs satisfactory?
3. Do the items of the Form-meaning Recognition Test of Academic Verbs discriminate satisfactorily between successful and less successful participants?
4. Should anymore items be reduced?
5. Are the distractors working satisfactorily?

Results

Starting with research question 1 descriptive statistics of the scores were calculated to get a first impression of how difficult the test had been for the students.³⁷ A Shapiro-Wilk test (Field, 2018, p. 251) showed that the data was normally distributed, $D(225)=0.862$, $p<0.001$, which allowed an interpretation of the mean (M) and the standard deviation (SD) (Dörnyei, 2007, p. 214). As visible in Table 7.13, students achieved a mean score of 68.59 points,

³⁴ For more detailed output see appendix 7.4.2.

³⁵ The Form-meaning Recognition Test of Academic Verbs used in the 1st data collection can be found in appendix 7.1.2 together with its solution.

³⁶ For more information on the data collection see chapter 6.2. For more information on participants or test administration see chapter 7.1.5.

³⁷ For more details on the descriptive statistics see appendix 7.4.3.

which is slightly lower than the median score in the piloting [$Mdn= 71.5$]. Individual scores deviated from the mean score by 8.60 points on average [$SD= 8.60$], which are 10% of the maximum score of 80 points. Hence, the group varies slightly but is quite homogeneous.

Value (s)	Total (80)
<i>M</i>	68.59
<i>SD</i>	8.60
α	0.86

Table 7.13: Descriptive results for Form-meaning Recognition Test of Academic Verbs 1st data collection (N=225)

To investigate test difficulty further facility values of all items were calculated. Table 7.14 presents a distribution of all test items using Bachman's (2004) categorization of test difficulty.

above 0.7	82.5%
0.3-0.7	17.5%
below 0.3	0%

Table 7.14: Distribution of facility values of Form-meaning Recognition Test of Academic Verbs (N=225)

Over 82% of all items were answered correctly by over 70% of the participants, while only 17.5% were between the range of 30 and 70 percent correct answers. While this seems at first sight not very reassuring, the results are convergent to what has been found in research on form-meaning recognition knowledge of vocabulary. Recall knowledge appears to be a greater burden for most L2 learners, while recognition knowledge is learned earlier and is therefore normally larger than the recall knowledge of the same words (e.g. Laufer, 2005; Schmitt, 2010, p. 21). Hence, facility values can be explained by prior research and correspond to expectations. However, in a next step it has to be shown that the test is reliable and that items discriminate well enough regardless of the scores clustered at the high end.

The reliability of the Form-meaning Recognition Test of Academic Verbs revealed that reliability was satisfactory (Field, 2018; Kline, 1999) with a Cronbach α of 0.90. However, as with the Form-meaning Recall Test of Academic Verbs this result is slightly below the reliability value of the piloting, namely $\alpha=0.96$. Referring to Henning's (1987) reasons for reduced scores of reliability measurement errors can be eliminated, since the scores were entered and double checked by myself and a second researcher. Nevertheless, similar to the Form-meaning Recall Test of Academic Verbs the administrative procedure has been altered from the piloting to the first data collection regarding the available time limit to fill in the two tests, which might be a potential factor for the change of Cronbach α score (Henning

1987, p. 77). Moreover, it might be possible that the piloting group was not comparable enough to the target group of the main study.

To gain further information item discrimination was investigated. Table 7.15 divides all 80 items of the Form-meaning Recognition Test of Academic Verbs into the four categories developed by Ebel (1979). Again, all items, which reach a score below 0.19 are considered to be poor, items between 0.20 to 0.29 marginal, items between 0.30 to 0.39 reasonable and items with a score above 0.40 very good.

	Number of items	Percentage
<i>very good</i>	6	7.50%
<i>good</i>	23	28.75%
<i>marginal</i>	20	25.00%
<i>poor</i>	31	38.75%

Table 7.15: Discrimination values Form-meaning Recognition Test of Academic Verbs (N=225)

As visible in Table 7.15 above, most items fall under the category of poor discrimination. However, as described the calculation of the discrimination value is sensible to the facility value of the item and hence a high number of correct or incorrect responses negatively influences the discrimination value of an item. Hence, the informative value of these numbers should not be overestimated in case of this test. Nevertheless, some items had clear negative discrimination values and were not beneficial to increase the overall test reliability, which is the reason that eight items were deleted in negotiation with the Form-meaning Recall Test of Academic Verbs. The eight items deleted were *restrict*, *focus on*, *rely*, *reject*, *decrease*, *deny*, *restore* and *abandon*. After the reduction of these eight items the discrimination values changed positively as can be seen in Table 7.16 below.

	Number of items	Percentage
<i>very good</i>	6	8.33%
<i>Good</i>	22	30.55%
<i>marginal</i>	19	26.38%
<i>Poor</i>	25	34.72%

Table 7.16: Discrimination values Form-meaning Recognition Test of Academic Verbs (N=225)

The reliability coefficient changed from $\alpha = 0.90$ to $\alpha = 0.89$, which means that it stayed stable.

7.2.4 2nd and 3rd data collection³⁸

As with the Form-meaning Recall Test of Academic Verbs the 2nd and 3rd data collection was not primarily focused on test development anymore and featured the same test formats comprising 72 academic verbs. For an account of test administration see the description of the administration of Form-meaning Recall Test of Academic Verbs. Considering the reliability of the Form-meaning Recognition Test of Academic Verbs Cronbach α stayed stable at 0.89, which can be considered satisfactory.³⁹

7.3 Collocation Recognition Test of Academic Verbs

The Collocation Recognition Test of Academic Verbs is a discrete, selective and context-independent test, since it tests vocabulary knowledge only, items were pre-selected by the researcher and no contextual knowledge is required to complete the test. Using Gyllstad's (2019, p. 389) categorization for multi-word unit tests, which is based on the categorizations by Laufer & Goldstein (2004) and Schmitt (2010), this test would be a form recognition test of verb + noun collocations.

7.3.1 Item selection

All target words included in the Collocation Recognition Test of Academic Verbs are taken from the original 95 academic verbs featuring in the Form-meaning Recognition and Recall Test of Academic Verbs. Regarding the type of collocation observed verb+noun combinations were selected as target units, since this type of collocation is not amongst the most frequent but tends to give the core meaning of messages (Gyllstad & Snoder, 2021).

25 items were chosen on the basis of the facility values of the target words in the first data collection of Form-meaning Recall Test of Academic Verbs, since it might be revealing to investigate whether any differences can be detected in collocation knowledge of familiar and not well-known words in future studies. Although mostly items with a low facility value (0.0-0.7) were considered, items with a high facility value (0.71-1) were additionally integrated (Bachman, 2004, p. 138). Table 7.17 presents the 25 items that were originally chosen for the Collocation Recognition Test of Academic Verbs categorized by their facility values.

³⁸ The Form-meaning Recognition Test of Academic Verbs used in the 2nd data collection can be found in appendix 7.1.3 together with its solution and the Form-meaning Recognition Test of Academic Verbs used in the 3rd data collection and its solution in appendix 7.1.4.

³⁹ See the SPSS output in appendix 7.4.4.

Facility value of target word in Form-meaning Recall Test of Academic Verbs1	0-0.29	0.3-0.4 0.6-0.7	0.4-0.6	0.7-0.8	0.8-1
	invoke (10) impose (4) amend (5) reinforce (8) reside in (2) promote (4)	attribute (4) pursue (5) assign (6) retain (3) attain (9) distribute (1)	maintain (2) predict (4) terminate (8) contribute (3) commit (4) induce (8)	approach (1) establish (1) perceive (2) shift (3) utilize (6)	assume (1) ensure (3)
Number of items	6	6	6	5	2

Table 7.17: Item selection for Collocation Recognition Test of Academic Verbs

Additional focus was put on a balanced distribution of frequency levels. Since the AWL (Coxhead, 2000) is divided into ten frequency bands each category of facility values includes a variety of words of different frequency. The frequency band of the individual target words is given in brackets next to the word in Table 7.17. Moreover, it was attempted to include as many items with multiple meanings as possible, since at least some the words of the Collocation Recognition Test of Academic Verbs should be used in the Multiple Meaning Recognition Test of Academic Verbs.

While 25 items seem to be a relatively small number in comparison to the target words included in other tests (Gyllstad, 2019), this test needs to be evaluated in its context of being used together with the other three tests developed covering the same academic verbs. Hence, while a larger number of target words and items might be desirable in general, including more collocations in this test might have resulted in overrepetition of the shared target words. Nonetheless, in the future more items could be developed for the Collocation Recognition Test of Academic Verbs.

7.3.2 Item development

With the help of the Corpus of Contemporary American English (COCA) collocations were generated for each of the 25 target words (Davies, 2008). The COCA is an electronic collection of authentic texts (Kennedy, 1998) encompassing 275,476 texts, which can be further divided into spoken texts, such as transcripts of interviews or TV programs, fictional texts as short stories or children stories, popular magazines, newspaper and academic journals. For the purpose of the development of this test we are interested in the section comprising academic journals and other texts prominent in academic discourse, since all collocations in the test should be specifically academic collocations (Davies, 2008). Hence,

all collocations included were generated from the 119,619,656 words of academic texts in the COCA.

Computerized corpora, such as the COCA, allow researchers to investigate native speaker use of the English language and to find language patterns unknown to them, such as collocations. The COCA allows searches for multiple word combinations, ranging from two-word combinations to language strings of ten interconnected words. Additionally, it is possible to indicate how many words are allowed to be between the node word and its collocate, which can be useful when searching for occurrences of for instance *build* and *house*, since the articles *the* or *a* might be found in between the two words. Moreover, it is possible to indicate the word class of the collocate searched for. For instance, it is possible to search for nouns only that collocate with the stimulus word *build*.

For the purpose of the Collocation Recognition Test of Academic Verbs only two-word combinations of noun+verb or verb+noun were searched for, where the verb was always one of the target words listed in Table 7.17. No words were allowed between the target verb and the noun.

For each of the 25 headwords the frequency of collocates and collocation strength in form of the MI-score was considered⁴⁰. On the basis of these numbers a minimum of three possible collocates were selected for each of the 25 headwords.

All distractors used in the Collocation Recognition Test of Academic Verbs were designed in accordance with Greidanus and Nienhus's (2001) suggestions. As their research shows, distractors which are semantically related to the correct answer function better than distractors with similarities in word form. Additionally, distractors with no relationship to the target word work similarly as semantically related distractors (Greidanus & Nienhus, 2001). Another rule of thumb adopted was Schmitt, Wun and Garras' (2011) finding that it is beneficial, if distractors are linked to each other, if possible. For instance, taking example (28) of Gyllstad's (2007) COLLEX 5, one can see that *receive a cold* and *achieve a cold*, both being distractors, seem to belong together somehow, since they are similar in form.

(28)

<i>receive a cold</i>	<i>achieve a cold</i>	<i>catch a cold</i>

All distractors developed based on these rules were checked in the COCA. Only collocates that did not appear at all in the corpus were considered as distractors for the test.

Additionally, all collocates, distractors or correct options, were checked for their frequency in the English language with the help of the program LexProfiler Compleat Web VP (Cobb, 2021), resulting in the following profile⁴¹:

⁴⁰ For an explanation and more detailed information about the MI- score see chapter 2.2.1 and appendix 7.5.1.

Frequency level	Cumulative token percentage
K-1	25.5%
K-2	55.9%
K-3	85.3%
K-4	94.1%
K-5	95.1%
K-7	98.0%
K-8	99.0%
K-10	100%

Table 7.18: Lexical profile of Collocation Recognition Test of Academic Verbs

According to this results 95% of all collocates are amongst the 5000 most frequent words of the English language, ensuring satisfactory understandability for the target group.

While several aspects were taken into consideration for the test design another valuable addition would have been the integration of learner corpora, such as the International corpus of Learner English (ICLE), which includes 5.7 million words of written English produced by adult university learners. A through analysis of the sub-corpus of L1 German speakers could provide additional insights into collocational use and behaviour of the target group and could be consulted for decisions regarding test development or scoring (Gyllstad & Snoder, 2021). For example, problem areas identified in the learners corpus, such as blending of two collocations or language correspondances might inform the choice of distractor items (Gyllstad & Snoder, 2021, p. 63). This would add an informed phraseological perspective to the test development additional to a rather statistical approach to test design (Gyllstad & Snoder, 2021, p. 68). However, this consideration postdates the development phase of the test instruments for this study and it remains unclear, which (sub)- corpus should be chosen for such an analysis, since this study is not solely composed out of L1 German speakers but had a minimum requirement of participants using German as an educational language.

Test design

After stimulus words, their potential academic collocates and distractors were selected, the test was developed aligning to two different test designs, namely the WAT (Read, 1993) and the COLLEX 5 (Gyllstad, 2007). The reason for not using a consistent test design throughout the data collection process, as for the Form-meaning Recall Test of Academic Verbs and the Form-meaning Recognition Test of Academic Verbs, was that in contrast to the two form-meaning link tests, the Collocation Recognition Test of Academic Verbs comprises only 25

⁴¹ For a more detailed lexical profile of the Collocation Recognition Test of Academic Verbs see appendix 7.2.3.

words. Hence, test effects might be more likely and substantial, since students could remember a large number of the test items regardless of a change in order. Therefore, students are given two different versions of the test at the two data collection points. However, the tests are only different in design but feature the same test items.

The first test format used was inspired by the Word Associates Test by Read (1993). While the original test mostly focuses on testing receptive association knowledge as well as receptive collocation knowledge as illustrated in example (29), the Collocation Recognition Test of Academic Verbs will focus only on testing receptive collocations.

Furthermore, in the original version Read (1993) presents the test takers with a headword in context of associates and collocates. As in example (29) the left column presents the stimulus word (*negative*) together with a group of potential associates (*pessimistic, surprised, bad*), while in the right column he lists potential collocates (*rhythm, attitude, union*) of the stimulus word. The correct answers are highlighted in italics.

(29)

Negative

<i>Pessimistic</i>	Surprised	<i>Bad</i>	Rhythm	<i>Attitude</i>	Union

Participants are required to select the correct associate/s and collocate/s. To avoid guessing, the number of correct answers is varied, which makes the scoring of the test relatively complex. Three different versions of scoring have been used so far, namely the one-point method (e.g. Nassaji, 2006; Qian & Schedl, 2004), the all-or-nothing method (Schoonen & Verhallen, 2008) or the correct-wrong method (Greidanus et al., 2004). While the one-point method awards one point for each correct answer, treating incorrect and missing answers the same, the all-or-nothing method gives only one point if all answers are correct (Schmitt, Wun & Garras 2011, p. 118). The correct-wrong method is more similar to the one-point-method awarding one point per correct answer but additionally giving one point per avoided distractor. Therefore, the correct-wrong method makes an attempt to distinguish between incorrect and no answer. Considering example (29) as model, the three different scoring methods would result in the following ranges of scores:

One-point method	0-3 points
All-or-nothing method	0-1 points
Correct-wrong method	0-6 points

Table 7.19: Distribution of points for different scoring methods of WAT

Schmitt, Wun and Garras (2011) investigated these three scoring methods, reaching the conclusion that all three methods were able to distinguish between different levels of ability and resulted in comparable results. However, they argue that the all-or-nothing method is more suitable if three answer options are given and the one-point-method is more

successful for four given collocates (Schmitt, Wun & Garras 2011, p. 119). Moreover, they state that the all-or-nothing method allows no conclusions about partial word knowledge, while the one-point-method considers partial knowledge at least to a certain extent (Schmitt, Wun & Garras 2011, p.122).

For the Collocation Recognition Test of Academic Verbs the test design of the WAT was adapted. Since this study is interested in receptive knowledge of collocations but not explicitly in receptive association knowledge, the left part of the column was eliminated. Instead, as in example (30), the 25 target words listed in Table 7.17 are presented in context of three possible collocates.

(30)

approach

retirement	☐	help	☐	sport	☐

As in Read's (1993) WAT format the number of correct responses varied from 1-3 in the first version. Convergent with Schmitt, Wun and Garras' (2011) findings the one-point method was used for scoring, since the study attempts to investigate also smaller changes in word knowledge, which makes consideration of partial word knowledge inevitable.

The second test format used as guidance was the COLLEX 5 by Gyllstad (2007). As can be seen in example (31) test takers, similarly to the WAT, are presented with three possible verb+noun collocates (Gyllstad 2007, p. 164). However, in the original version of COLLEX 5 only one answer was correct and two options were carefully designed distractors. Hence, the scoring method of the COLLEX 5 is more straightforward with 1 point awarded, if the correct answer was ticked and 0 points, if one of the distractors was chosen.

(31)

receive a cold *achieve a cold* *catch a cold*

Building on previous knowledge about the probability of guessing in similar test formats, Gyllstad (2007, p.75) decided on incorporating a self-report system to estimate the amount of guessing that takes place when taking COLLEX 5. More precisely students were asked to indicate for each item if they were sure about their answer, or if they had to guess. While being aware that self-report data about guessing behavior is not completely reliable for various reasons, one can argue that students who were taking Gyllstad's (2007) test in the piloting phase had no reason to be dishonest regarding their guessing behavior, since the results had no direct impact on them.

For the second version of the Collocation Recognition Test of Academic Verbs the COLLEX 5 format was adapted slightly. As visible in example (32) test takers were presented with three possible two-word combinations similar to the original format. However, instead of one correct option per item, there could be up to three correct collocates per item.

(32)

a) approach retirement b) approach help c) approach sport

a	b	c
☐	☐	☐

correct answer/s

Hence, the same scoring procedure as described above for the WAT above was adopted, giving 1 point per correct answer. Moreover, the same self-report procedure as described by Gyllstad (2007) was adopted for both test formats of the Collocation Recognition Test of Academic Verbs. Example (33) illustrates how guessing was integrated in the WAT format.

(33)

E.1. make

shadow	☐	room	☒	help	☐

This person thinks that room is the correct answer but he/she had to guess.

In the WAT version of the Collocation Recognition Test of Academic Verbs participants were asked to tick the box of the answer/s they thought were correct. In a second step they were required to circle the word if they had been guessing. Hence, in example (33) the person thinks that *make shadow* is the correct answer but he or she had been guessing and did not know the answer.

In the COLLEX format of the Collocation Recognition Test of Academic Verbs participants were asked to tick the answers that they thought were correct in the column labeled *correct answer/s*. Additionally, they had to indicate in the row below, which is labeled *need of guessing*, whether they had been guessing for the answer or not. Example (34) illustrates this procedure.

(34)

a) make shadow

b) make room

c) make help

✖	✖	
✖		

correct answer/s

need of guessing

In case of this example the test taker thinks that *make room* is a correct answer and additionally guesses that *make shadow* could be a possible correct answer.

7.3.3 Piloting 1⁴²

The first piloting of the two formats of the Collocation Recognition Test for Academic Verbs was conducted in May with 19 students of the English Department of the University of Vienna⁴³. All students were taking part in a tutorial for *History of English*, which is a lecture in the 2nd to 3rd semester of the study program. Beforehand all items were checked by a

⁴² The Collocation Recognition Test of Academic Verbs used for the 1st piloting can be found in appendix 7.1.11 together with its solution. Moreover, the SPSS output for the COLLEX piloting results can be found in appendix 7.5.2.1 (traditional one-point scoring) and 7.5.2.2 (partial one-point scoring), of the WAF in appendix 7.5.2.3 (traditional one-point scoring) and 7.5.2.4 (partial-one point scoring), and the results regarding guessing behaviour in appendix 7.5.2.5.

⁴³ For more details insights into the data collection procedure see chapter 6.2.

native speaker of English, who gave feedback on the instructions and the test format as well as on individual items.

Research questions

In the first piloting the following questions were attempted to be answered:

1. Is the Collocation Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Do the distractors work satisfactorily?
3. Is the Collocation Recognition Test of Academic Verbs reliable?
4. Does the Collocation Recognition Test of Academic Verbs discriminate well enough between good and less successful students?
5. Is the Collocation Recognition Test of Academic Verbs prone to guessing?
6. Is there a significant correlation between the two different test formats?
7. Can any patterns regarding guessing behavior be detected?

Test administration

To avoid bias by order of reception, half of the participants were first given the WAF format of the Collocation Recognition Test of Academic Verbs and the other half was first given the COLLEX format. When they were done with the first version, participants were asked to raise their hands and they were given the second test format, while the filled in test was collected to avoid cheating. Participants were informed why the tests contained the same items only after they had finished the tests. During the whole timespan students had the opportunity to ask questions and the researcher was present the whole time. Students took on average 20 minutes to fill in both tests. There was no time limit given, which resulted in many students being finished early but few students taking their time. This should be considered in future data collections.

Results

In the following the results are presented separately for the two different test formats. Since the two tests revealed slightly different results but had to cover the same items, adaptations made to the Collocation Recognition Test of Academic Verbs was oriented only on the COLLEX format, since it was the more reliable format. Moreover, two different scoring variants of the one-point method have been put to the test. In both scoring methods 1 point is rewarded for a correct answer. However, while in the traditional one-point method the headwords are the focus, the partial one-point method looks at individual items. To illustrate this, example (35) would be scored with 1 point out of 2 for the headword *establish* with the traditional one-point method.

(35)

2. a) establish advice b) *establish rapport* c) *establish trust*

a	b	c
✗		✗

correct answer/s
need of guessing

Correct answers are highlighted in italics for illustration. If example (35) were scored with the partial one-point method, *establish rapport* would receive 0 points out of 1 and *establish trust* would receive 1 point out of 1. However, in contrast to the traditional one-point method, which is interested in the headword *establish*, the partial one-point method treats the individual answers (*establish rapport* and *establish trust*) independently of their headword. This difference in the values used when coding and working with the data are not irrelevant for reliability as will be seen in the individual results.

COLLEX

a) Partial One-point scoring

Despite the small sample number, the data of the COLLEX version of the Collocation Recognition Test of Academic Verbs was normally distributed when using the partial one-point scoring method, which gives credit for partial word knowledge [$D(19)=0.932$, $p=0.188$]. The mean score reached by the piloting group using the partial one-point scoring method was $M=30.42$ points out of 51 points. The standard deviation of $SD=5.40$ equates a variation of 10.59% of the total score⁴⁴.

Value (s)	Total (51)
<i>M</i>	30.42
<i>SD</i>	5.40
<i>α</i>	0.67

Table 7.20: Descriptive results COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)

Item facility was quite sufficient with 47.06% of all items being between a facility value of 0.3 and 0.7 (Bachman, 2004, p. 138). As Table 7.21 shows only 15.69% were below 0.3 and 37.25% above 0.7, being relatively easy for participants.

above 0.7	37.25%
0.3-0.7	47.06%
below 0.3	15.69%

Table 7.21: Distribution of facility values COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)

⁴⁴ For more detailed information on the descriptive statistics see appendix 7.5.2.2.

In a next step the distractors were analyzed to check whether they functioned sufficiently. Anderson and Morgan (2008, p. 79) state that a distractor is satisfactory if it is chosen by at least five percent of all participants. However, as with most values boundaries for acceptance are a topic of debate, which is why a slightly stricter threshold of at least ten percent of all participants has been decided on.

Five out of 24 distractors reached a score below ten percent, namely *advice*, *spring*, *memories*, *birthday* and *day*. These distractors were all replaced in the next version of the Collocation Recognition Test of Academic Verbs. Additionally, *help*, *heaven*, *content*, *steps*, *chance* and *time* were deleted due to them being possible answers according to feedback by native speakers or due to the fact that they were working better than the actual answer, making the item too difficult for the intended audience.

The reliability of the COLLEX format of the Collocation Recognition Test of Academic Verbs was at first sight not fully satisfactory, reaching a Cronbach's α of only 0.67. Hence, improvements had to be made by looking at reliability values with the individual items as well as discrimination values.

When looking at item discrimination one can see that convergent to reliability item discrimination tends to be rather poor. Of 51 items in total, two items were excluded from the analysis, since they had zero variance, which is the reason that Table 7.22 reports on the discrimination results of 49 items.

	Number of items	Percentage
<i>very good</i>	7	14.29%
<i>Good</i>	7	14.29%
<i>marginal</i>	7	14.29%
<i>Poor</i>	28	57.14%

Table 7.22: Discrimination values COLLEX format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)

An equal amount of 14.29% falls on very good, good and marginal items, while a totality of 57.14% of all items are reported to have poor discrimination.

b) Traditional one-point scoring

Since both scoring methods result in the same sum index, descriptive statistics stay the same. However, facility values get more multi-layered when looking at the traditional one-point scoring method. While with the partial one-point scoring method all items seemed to be comparable, this method shows that the items are not all similar due to the different number of correct answers per item. While one stimulus word has only one correct answer and hence a maximum of 1 point to be reached, others have two or even three correct

options. In the following only items with the same number of correct answers are compared.

As Table 7.23 shows that 18 items contained two correct answers, while only four items had three correct items and three items had only one correct answer⁴⁵.

Regarding items with two correct answers one can see that on average 8.48% of the participants knew no answer at all. For completeness it has to be mentioned that *reside in* was excluded from this calculation, since it was a clear outlier with 52.6% incorrect answers. On average 63.74% of all participants knew at least one answer of two, which leaves 27.78% of participants who ticked both answers correctly.

	Number of items	All incorrect	1 correct answer	2 correct answers	3 correct answers
One correct answer	3	35.1%	64.9%		
Two correct answers	18	8.48%	63.74%	27.78%	
Three correct answers	4	0%	27.63%	40.85%	31.52%

Table 7.23: Item facility with traditional one-point scoring COLLEX Collocation Recognition Test piloting 1 (N=19)

Items with only one correct answer were answered correctly by 64.9% of all participants, which is comparable to the number of people who knew at least one correct answer of items with multiple correct answers. For items with three correct answers only 27.63% of all participants ticked only one answer, while 40.85% decided on two correct options.

So far, no large differences can be detected between the traditional one-point scoring and the partial one-point scoring. However, an analysis of reliability shows that with the traditional one-point scoring the Cronbach α rises to 0.73, while for the partial one-point scoring it was only 0.67. Hence, the traditional one-point scoring appears to foster reliability of the test.

Regarding item discrimination Table 7.24 shows that while the item discrimination is also not perfect for the traditional one-point scoring, it is definitely better than for the partial one-point method.

⁴⁵ For more details see appendix 7.5.2.1.

	Number of items	Percentage
<i>very good</i>	8	32.0%
<i>Good</i>	5	20.0%
<i>Marginal</i>	5	20.0%
<i>poor</i>	6	24.0%

Table 7.24: Discrimination values with traditional one-point scoring COLLEX Collocation Test piloting 1 (N=19)

Hence, it seems that the traditional one-point scoring method, which looks at each item individually is preferable to the partial one-point scoring method. Moreover, all six items with poor discrimination value would increase the Cronbach α if they were to be deleted. The items in question are *approach*, *establish*, *attribute*, *invoke*, *assume* and *perceive*.

Based on these values a second version of the Collocation Recognition Test of Academic Verbs was developed. However, it was decided that the new version should no longer be a multiple-choice test but was changed to a single choice test. The reason was that the items are simply not comparable regarding difficulty, if they do not have the same number of correct items. Furthermore, for most items discrimination or facility value of one correct answer was not satisfactory, which was the reason why for each item one correct answer with the best discrimination and facility value was kept. To substitute the deleted correct answers additional distractors were created. All statistical decisions reported were taken on the basis of expert advice⁴⁶.

WAF

a) Partial one-point scoring

Similar to the COLLEX format scores were normally distributed, $D(19)=0.355$, $p=0.355$. The mean score reached was $M=31.28$ out of 51 points, which is slightly more than for the COLLEX format⁴⁷. The standard deviation was 4.46, which is slightly lower than for the COLLEX format. Hence, the group presents itself as more homogenous in the WAF format.

Value (s)	Total (51)
<i>M</i>	31.28
<i>SD</i>	4.46
α	0.54

Table 7.25: Descriptive results WAF format Collocation Recognition Test of Academic Verbs piloting 1 (N=19)

⁴⁶ Thanks to Jakob Smigierski, M.Sc.

⁴⁷ For more information see appendix 7.5.2.4.

Item facility was comparable to the COLLEX format. However, slight differences were found as can be seen in Table 7.26.

above 0.7	41.18%
0.3-0.7	43.14%
below 0.3	15.69%

Table 7.26: Item facility with partial one-point scoring WAF format piloting 1 (N=19)

To locate the items that differed the most from each other, facility values were compared. For eight items the difference between the COLLEX format and the WAF format was great enough for them to change the category of facility values. These items were *maintain contact*, *attribute success*, *pursue careers*, *distribute funds*, *induce changes*, *perceive sound*, *shift attention*, *shift responsibility*.

Regarding distractors similarly to the COLLEX format *spring* and *birthday* were chosen by less than ten percent of the participants and *help* and *content* made the item too difficult for the intended audience. Hence, there is a wide agreement between the two formats concerning the quality of distractors.

Divergent to item facility similarities can be detected in item discrimination between the COLLEX and the WAF format. As can be seen in Table 7.27 only twelve percent of all items could be assigned to the categories very good, good and marginal, while 64% of all items were poorly distinguishing between good and less successful learners.

	Number of items	Percentage
<i>very good</i>	6	12.0%
<i>good</i>	6	12.0%
<i>marginal</i>	6	12.0%
<i>poor</i>	32	64.0%

Table 7.27: Discrimination values partial one-point scoring WAF format piloting 1 (N=19)

As in the COLLEX format this rather poor result goes hand in hand with a rather low reliability coefficient of $\alpha=0.54$. Out of the 32 items with poor discrimination value, 21 would increase the reliability coefficient, if they were deleted.

b) Traditional one-point scoring

A detailed analysis of item facility with the traditional one-point scoring in the WAF test resulted in similar findings as in the traditional one-point scoring of the COLLEX format. Table 7.28 gives the percentages of correct answers for the different items for the WAF test. For easier comparison the numbers of the COLLEX format can be found in brackets beneath the results of the WAF format.

	Number of items	All incorrect	1 correct answer	2 correct answers	3 correct answers
<i>One correct answer</i>	3	27.77% (35.1%)	72.23% (64.9%)		
<i>Two correct answers</i>	18	9.89% (8.48%)	65.43% (63.74%)	24.68% (27.78%)	
<i>Three correct answers</i>	4	0% (0%)	25.0% (27.63%)	38.88% (40.85%)	36.12% (31.52%)

Table 7.28: Item facility with traditional one-point scoring collocation WAF format piloting 1 (N=19)

Divergent to the difference in reliability score between partial and traditional one-point scoring in the COLLEX format, reliability falls slightly from to $\alpha=0.48$.⁴⁸

Regarding item discrimination the values were exactly the same as for the COLLEX format as can be seen in Table 7.29.

	Number of items	Percentage
<i>very good</i>	8	32.0%
<i>good</i>	5	20.0%
<i>marginal</i>	5	20.0%
<i>poor</i>	6	24.0%

Table 7.29: Discrimination values with traditional one-point scoring WAF format piloting 1 (N=19)

Hence, similar to the COLLEX format the WAF format appears to function better with the traditional one-point method of scoring. However, overall the WAF format achieved less satisfactory results than the COLLEX format, which is the reason why the COLLEX format was taken as orientation point for decisions about improvements.

Despite the differences on item level between the COLLEX and the WAF format a Spearman's Rho correlation (Field, 2018, p. 351) shows that there is a significant, strong positive correlation between the COLLEX and the WAF format of the Collocation Recognition Test of Academic Verbs, $r_s=0.735$, $p<0.001$ ⁴⁹. Moreover, a Kendall's tau correlation was computed. While the Kendall's tau correlation is less established, Field (2018, p.353) states that it is the better estimate of a population, if working with a small data set with a large

⁴⁸ While this result is not satisfactory, it has to be considered that with the traditional one-point scoring reliability of 25 items is evaluated. In contrast, the partial one-point scoring comprises 49 items. This difference might influence the reliability value slightly, since Cronbach alpha is influenced by item number.

⁴⁹ This correlation was computed for the two traditional one-point scoring approaches of COLLEX and WAF format. The results of the correlation between the two formats with the partial one-point scoring can be found in appendix 7.5.2.5.

number of tied ranks. Similar to the Spearman's Rho correlation, the Kendall's tau correlation suggests that there is a significant, strong positive correlation between the two formats, $\tau=0.593$, $p<0.001$. However, the correlation is slightly weaker, which could be due to the fact that there were not many tied ranks in the data set at hand. Hence, the Spearman's Rho correlation might be even more suitable in this case. On the basis of these results one can conclude that the two formats of the Collocation Recognition Test of Academic Verbs are strongly related and can be used interchangeably at the two data collection points.

Lastly, the seventh research question regarding guessing behavior will be attempted to be answered. For the COLLEX format the number of guesses overall ranged from 0 to 20⁵⁰. On average 8.94 guesses were made by the participants, [$D(16)=0.937$, $p=0.317$]. However, the standard deviation of 5.98 suggests that the group was quite diverse regarding the number of guesses reported. Out of all guesses made 0-14 were correct. The mean number of correct guesses for the COLLEX format was 6.69 with a standard deviation of 4.38, [$D(16)=0.961$, $p=0.676$]. Hence, the group is again rather heterogenous. On the other hand, 0-8 incorrect guesses were made. The median number of incorrect guesses was 1 with a Q of 2.

In a next step the participants were divided into three groups according to their performance on the Collocation Recognition Test of Academic Verbs. Participants who reached a score above 33 were considered to be highly proficient (N=6), participants with a score between 26-32 mid-proficient (N=6) and participants with a score less than 26 points low-proficient (N=4). Since the numbers of the individual groups are very small, comparison was restricted to descriptive statistics.

As Table 7.30 shows there is no clear decrease of guessing with increased proficiency. While the low proficiency group appeared to make the most guesses overall, the high-proficiency group made the next highest attempts to guess and the mid-proficiency group reported to guess the least.

	High- proficiency group (N=6)	Mid- proficiency group (N=6)	Low proficiency group (N=4)
<i>Number of guesses</i>	Range: 2-20 <i>M</i> =8.71	Range: 0-14 <i>M</i> =5.6	Range: 5-20 <i>M</i> =12.0
<i>Number of correct guesses</i>	Range: 1-14 <i>M</i> =6.71	Range: 0-13 <i>M</i> =6.8	Range: 3-12 <i>M</i> =8.0

⁵⁰ Only the results for the COLLEX format are given, since guessing behaviour on the two test formats was nearly identical.

<i>Number of incorrect guesses</i>	Range: 1-6 <i>M</i> =2.0	Range: 0-4 <i>M</i> =1.2	Range: 1-8 <i>M</i> =3.75

Table 7.30: Guessing behavior sorted by proficiency Collocation Recognition Test of Academic Verbs piloting 1 (N=16)

Moreover, it is surprising that in regard to correct guesses the low-proficiency group guessed most correctly ($M=8.0$), followed by the mid-proficiency group ($M=6.8$) and only in the last place are the high-proficiency participants ($M=6.71$). It will have to be seen in further studies whether this rather unusual picture of guessing behavior is typical for the Collocation Recognition Test of Academic Verbs and requires adaption.

7.3.4 Piloting 2⁵¹

On basis of the not fully satisfactory results of the first piloting a second piloting phase of the two formats of the Collocation Recognition Test of Academic Verbs was conducted in late May⁵². In this piloting the adapted single choice version of the COLLEX and the WAF format was administered.

Research questions

Similar to the first piloting round the research questions under investigation were:

1. Is the Collocation Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Do the distractors work satisfactorily?
3. Is the Collocation Recognition Test of Academic Verbs reliable?
4. Does the Collocation Recognition Test of Academic Verbs discriminate well enough between good and less successful students?
5. Is the Collocation Recognition Test of Academic Verbs prone to guessing?
6. Can any patterns regarding guessing behavior be detected?

Participants

28 participants were included in the second round of piloting. Out of the total number, 17 participants were part of a voluntary tutorial for an introductory linguistics lecture and eleven participants were English students who volunteered to participate outside of a course setting. However, all students were in their first semesters of their studies and did not complete the language competence courses in the focus of this study.

⁵¹ The Collocation Recognition Test of Academic Verbs used for the 2nd piloting can be found in appendix 7.1.12 together with its solution. Moreover, the SPSS output of the COLLEX piloting results can be found in appendix 7.5.3.1 and of the WAF in appendix 7.5.3.2.

⁵² For more detailed information about the data collection process have a look at chapter 6.2.

Test administration

Half of the whole participant group received the updated single choice COLLEX format first, followed by the updated WAF format. To avoid bias through order of test completion the other half filled in the WAF format before the COLLEX format. All participants had the opportunity to ask questions and the researcher was present the whole time.

Results

Similar to the first round of piloting the results are going to be reported separately for the two different formats of the Collocation Recognition Test of Academic Verbs. However, since the format has been changed from a multiple choice to a single choice test, there is only one scoring method. Participants received 1 point on ticking the correct collocation and 0 points, if they ticked nothing or a distractor. Since the test comprised 25 items a maximum score of 25 points was possible.

COLLEX

For the single choice COLLEX format the data was normally distributed, [$D(28)=0.944$, $p=0.138$]. On average participants reached $M=16.36$ points out of 25, which equates to 64% of the total score. Hence, the test appears to be of mid difficulty for the participants. However, the standard deviation was 3.96 points, which means that on average participants varied from the mean by 15.84%. Therefore, the group was not entirely homogeneous. The scores ranged from a minimum of 7 points to a maximum of 22 points. Table 7.31 summarizes the most important descriptive findings.

Value (s)	Total (25)
<i>M</i>	16.36
<i>SD</i>	3.96
<i>α</i>	0.71

Table 7.31: Descriptive results COLLEX format piloting 2 (N=28)

Regarding facility values 52% of all items reached a score between 0.3-0.7, 4% of all items were too difficult reaching a score below 0.3 and 44% of the items were too easy with a facility value above 0.7 (see Table 7.32).

above 0.7	44.0%
0.3-0.7	52%
below 0.3	4%

Table 7.32: Item facility COLLEX format piloting 2 (N=28)

Similar to the stimulus words all distractors were checked for their facility value. 37 out of 50 distractors were chosen by more than 10% of the participants, which equates 74%. 13 distractors were chosen by fewer than 10%. As some borderline cases were still kept for the

next version of the test five distractors (*agenda, solicitors, deployment, concern, lies*) were deleted.

Since the Collocation Recognition Test of Academic Verbs still included too many easy items, several stimulus words were deleted and exchanged for new target words. However, the decision which items were substituted was not only based on the facility value of items but additionally on the discrimination values and reliability score.

An analysis of the discrimination values of all items resulted in 72% of all items having an at least satisfactory discrimination. The exact distribution can be found in Table 7.33 below.

	Number of items	Percentage
<i>very good</i>	7	28.0%
<i>Good</i>	7	28.0%
<i>Marginal</i>	4	16.0%
<i>Poor</i>	7	28.0%

Table 7.33: Discrimination values COLLEX format piloting 2 (N=28)

Additionally, it was decided that six out of the seven remaining items with poor discrimination were kept, since they were part of the Multiple Meaning Recognition Test, which should have at least partially overlapping items. Moreover, the number of participants in this piloting is still too small to take grave decisions. Overall, four items were deleted from the test, namely *reinforce, perceive, shift and utilize*. All four words were too easy for participants with facility values above 0.85. In exchange five new stimulus words were incorporated in the test. Items for *infer, allocate, exceed, inhibit* and *enhance* were created.

The reliability coefficient of the test increased to $\alpha=0.71$ in comparison to piloting 1, which can be considered to be satisfactory for a test that covers 25 items (Field, 2018, p. 823; Kline, 1999).

WAF

The data for the single choice WAF format was not normally distributed, [$D(27)=0.887$, $p=0.007$]. Therefore, the median and the interquartile range are taken as reference points in descriptive data analysis (Bachman, 2004, p. 77). The median for the WAF format was relatively similar to the COLLEX format, the previous being 16.37 and the latter 17 (see Table 7.34). However, Q is even higher with a variation of 5 points on average from the median, than the SD in the COLLEX format. Hence, participants varied even more strongly in their performance when using the WAF format of the Collocation Recognition Test of Academic Verbs. The results ranged from a minimum of 7 to a maximum of 23.

Value (s)	Total (25)
<i>MD</i>	17.0
<i>Q</i>	5
α	0.70

Table 7.34: Descriptive results WAF format piloting 2 (N=28)

Concerning facility values the picture is mostly similar as with the COLLEX format. Again 52% are in the optimal range of 0.3- 0.7, 8% of all items are too difficult and 40% are too easy (see Table 7.35).

<i>above 0.7</i>	40.0%
<i>0.3-0.7</i>	52%
<i>below 0.3</i>	8%

Table 7.35: Item facility WAF format piloting 2 (N=28)

For six items the difference between the facility value in the WAF format and the COLLEX format is large enough that they were assigned to a different category. Those items were *attribute*, *pursue*, *invoke*, *impose*, *assign* and *commit*. The more so as those cases were borderline cases between two categories in the COLLEX format already.

While the facility values were rather similar in the two test formats, distractor analysis resulted in different findings. In the WAF format only 29 distractors were above the threshold of 10% and 21 distractors were below it. Nevertheless, the values were a bit more evenly distributed with few clearly too low distractors. Distractors, which were only slightly below the 10% border in one of the formats were kept. However, as already stated above five distractors had to be deleted.

Overall the analysis of discrimination in the WAF format was slightly worse than in the COLLEX format with 68% of all items having at least satisfactory discrimination. However, the discrimination values of the WAF format confirm that the decision to keep the four items, which are found in the Multiple Meaning Recognition Test was correct. All four items display better discrimination values in the WAF format. Hence, it seems to be too early to draw final conclusions whether they should be excluded from the test in general or not.

Reliability is similar for the WAF format as for the COLLEX format, with $\alpha=0.70$. Therefore, adaptations made appear to be fruitful.

7.3.5 2nd data collection⁵³

Since the Collocation Recognition Test of Academic Verbs and the Multiple Meaning Recognition Test were first used in the 2nd data collection, there are no results for the first data collection. In the 2nd data collection the WAF format of the Collocation Recognition Test of Academic Verbs was given to the students⁵⁴.

Research questions

Since the decision was not to make certain changes in the piloting phase due to the small sample number, the Collocation Recognition Test is again being investigated for the following research questions:

1. Is the Collocation Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Are the distractors working satisfactorily?
3. Is the Collocation Recognition Test of Academic Verbs reliable?
4. Is the Collocation Recognition Test of Academic Verbs discriminating well enough between good and less successful students?
5. Is the Collocation Recognition Test of Academic Verbs prone to guessing?
6. Can any patterns regarding guessing behavior be detected?

Results

The scores of the second data collection were not normally distributed, [$D(202)=0.985$, $p=0.029$]. The median score reached was 16 points out of 26, which equates 61.54%. Hence, the test seemed to be of mid difficulty for the participants. The interquartile range was 6 points, which equates 23.08% of the maximum score. Therefore, test performance seemed to be not completely homogenous, which can additionally be supported by the range covering 6-24 points.

Value (s)	Total (26)
<i>MD</i>	16
<i>Q</i>	6
<i>α</i>	0.64

Table 7.36: Descriptive results Collocation Recognition Test of Academic Verbs 2nd data collection (N=202)

⁵³ The Collocation Recognition Tests of Academic Verbs used for the 2nd and 3rd data collection can be found in appendix 7.1.13 and 7.1.14 together with their solutions. Moreover, the SPSS output for the analysis of the 2nd data collection can be found in appendix 7.5.4.

⁵⁴ For more details on the participants and test administration see one of the descriptions of the 2nd data collection for Form-meaning Recall Test of Academic Verbs or Form-meaning Recognition Test of Academic Verbs.

Regarding facility values a similar picture as in the piloting rounds was found. As Table 7.37 illustrates 53.85% of all items had a satisfactory facility value, while 5.85% of the items were too difficult and 42.31% were too easy.

above 0.7	42.31%
0.3-0.7	53.85%
below 0.3	3.85%

Table 7.37: Item facility Collocation Recognition Test of Academic Verbs 2nd data collection (N=202)

Distractor analysis showed that 32 out of 52 distractors were satisfactory, reaching a level above 10%. 18 distractors were used by less than 10% of the participants. However, since distractors are working as pairs against the correct option, the facility value was still considered satisfactory, if at least one distractor of the pair was working well.

A reliability analysis showed that the Cronbach α was lower with $\alpha=0.64$. Closer investigation of the discrimination values showed that nine items should be deleted, namely *approach*, *predict*, *pursue*, *invoke*, *contribute*, *reside*, *enhance*, *promote* and *exceed*. However, most of the items, such as *approach*, *promote* and *pursue* could not be deleted due to their inclusion in the Multiple Meaning Recognition Test of Academic Verbs.

7.4 Multiple Meaning Recognition Test of Academic Verbs

The format of the Multiple Meaning Recognition Test of Academic Verbs was inspired by several previous tests. Tannenbaum et al. (2006) based his test of multiple meanings on the LPT_R measure by Richard and Hanner (1995). In this multiple meaning test participants were presented with the target word in three correct but different sentence contexts. For each of the different applications of the target word they had to either name a definition or a near synonym (Tannenbaum et al., 2006). Li and Kirby (2015) adapted this multiple meaning test by presenting the target word in four different sentence contexts. However, not all uses of the target word were correct and the participants had to indicate, in which sentences the word was used correctly. The correct answers varied from 0 to 4, which means that the test was a multiple-choice test and participants had no information how many items were correct or incorrect (Li & Kirby, 2015). The test format closest to the Multiple Meaning Recognition Test of Academic Verbs is the test developed by Fernández-González and Schmitt (2019), which was developed simultaneously to this multiple meaning recognition test⁵⁵. Similar to Li and Kirby (2015) they present the target word in different sentence contexts. However, they have five possible sentential uses of the target word instead of the four possible options provided here. Moreover, their participants were

⁵⁵ While the two tests were developed at the same time insights into González-Fernández and Schmitts' (2019) multiple meaning recognition test was possible at the BAAL VOCAB SIG 2019 in Manchester.

informed that there is a maximum of three correct answers and hence a minimum of two distractors.

7.4.1 Item selection

In the Multiple Meaning Recognition Test of Academic Verbs twelve of the original target words with multiple meanings are employed. The words chosen are *approach*, *establish*, *maintain*, *pursue*, *impose*, *assume*, *affect*, *perceive*, *commit*, *attribute*, *promote* and *conceive*. Similar to the Form-meaning Recognition Test of Academic Verbs the Cambridge English Dictionary (University of Cambridge, n.d.), the Oxford English Dictionary (Oxford University Press, n.d.), the Oxford Learners Dictionaries (Oxford University Press, n.d.) and Linguee (DeepL GmbH, n.d.) were used as sources to locate potential meanings. The primary meaning, which is included in the Form-meaning Recognition Test and the Form-meaning Recall Test of Academic Verbs, is included as one potential sentence context. A secondary meaning was chosen on the basis of the frequency of results in the four dictionaries. The secondary meaning, which was most prominent in all four dictionaries was included in the Multiple Meaning Recognition Test. For instance, the primary meanings for *establish* found in the four dictionaries were:

- to found; to set up (eg. a university, a business) (Cambridge English Dictionary)
- establish something to start or create an organization, a system, etc. that is meant to last for a long time (Oxford English Dictionary)
- set up on a firm or permanent basis. (Oxford Learners Dictionaries)
- etwas aufbauen (Linguee)

While the most prominent secondary meaning for *establish* found in all four dictionaries was:

- to show to be true; to prove (Cambridge English Dictionary)
- to discover or prove the facts of a situation (Oxford English Dictionary)
- show (something) to be true or certain by determining the facts. (Oxford Learners Dictionaries)
- bestimmen (Linguee)

7.4.2 Item development

After the primary and the secondary meaning of the target words had been decided on, items were developed. Two sentences presented the target word once with the correct primary and once with the correct secondary meaning. Example (36) illustrates the format with the example of *establish*. Sentence a) would be the correct use of the primary meaning of *establish*, while sentence c) correctly uses the secondary meaning of *establish*.

(36)

establish

a) He <i>established</i> an aspiring company in a short time.	a) ☐
b) I forgot to <i>establish</i> the tailgate.	b) ☐
c) The doctors were able to <i>establish</i> the age of the corpse.	c) ☐
d) No party is allowed to <i>establish</i> its rights to third parties.	d) ☐

Options b) and d) are distractors, which have been developed on the basis of the principles described in Greidanus and Nienhus (2001), and Schmitt, Wun and Garras (2011). It was attempted to create distractor sentences that would be correct, if a semantically related word to the target item was used. In sentence d) the word *impose* was taken as inspiration, with the correct sentence being *No party is allowed to impose its rights on third parties*. Additionally, the preposition has been altered to ensure challenge participants.

In contrast to example (36) example (37) has no relationship between the original word and the target word:

(37) perceive

After the interruption, the professor *perceived* with her lecture.

According to Greidanus and Nienhus (2001) this is the second most effective way to develop distractors. Very rarely form-based distractors were developed as can be seen in example (38):

(38) affect

- a) I do hope that my honest apology *affects* him.
- b) The protest *effected* a change.

While sentence a) uses the target word *affect* correctly, sentence b) uses the word *effect*, which is very similar in form and can easily be mistaken.

For each target item two correct sentence contexts and two distractors were developed. Similar to the other tests a lexical profile of the sentence context was made with LexProfiler VP-Compleat (Cobb, 2021)⁵⁶. Table 7.38 illustrates clearly that the sentential context in which the target words are embedded should be understandable for the target group of participants, since over 95% of all words can be found in the 3000 most frequent words of the English language.

⁵⁶ A more detailed lexical profile of the Multiple Meaning Recognition Test of Academic Verbs can be found in appendix 7.2.4.

Frequency level	Cumulative token percentage
K-1	81.9%
K-2	88.9%
K-3	96.7%
K-4	97.2%
K-6	97.5%
K-9	97.7%
Off list	100%

Table 7.38: Lexical profile of Multiple Meaning Recognition Test of Academic Verbs

Participants were informed in the instructions as well as orally at the sittings that two options were correct, since this made scoring considerably easier. There were two possible options for scoring. Either one point was given for the correct identification of both possible meanings of the target word and 0 points were given, if one or two answers were incorrect. This scoring is called headword scoring in the following. The other method of scoring would be partial credit scoring with 1 point awarded for each correct option chosen, which means that each target word is worth 2 points maximum.

7.4.3 Piloting⁵⁷

The Multiple Meaning Recognition Test of Academic Verbs was first taken and reviewed by two native speakers of English, as well as colleagues and then piloted in May as visible in chapter 6.2.

Research questions

1. Is the Multiple Meaning Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Does the Multiple Meaning Recognition Test of Academic Verbs discriminate well enough between successful and less successful students?
3. Do the distractors of the Multiple Meaning Recognition Test of Academic Verbs work well enough?
4. Is the Multiple Meaning Recognition Test of Academic Verbs reliable?

Participants

The test was taken by 23 students taking the Tutorial of the *History of English* lecture course.

⁵⁷ The Multiple Meaning Recognition Test of Academic Verbs used in the piloting can be found in appendix 7.1.16 together with its solution and the SPSS output of the partial credit scoring in appendix 7.6.1.2 and the headword scoring in appendix 7.6.1.1.

Test administration

Participants were asked to fill in the Multiple Meaning Recognition Test of Academic Verbs silently and independently. Before starting participants were given a form of consent and were informed verbally about the purpose of the data collection. Participants had enough time to ask questions. The researcher was present the whole time during data collection. Participants took about eight minutes to complete the test.

Results

The results will be split up into the results for the headword scoring method and the partial credit scoring method.

Headword scoring method

The data collected was normally distributed, [$D(23)=0.919$, $p=0.062$]. However, since the result was only very weakly normally distributed both the mean and the median will be interpreted. Of a maximum score of 12 points participants reached a mean score of 6.87. Therefore, the participants were able to tick both correct answers for about half of all target items. The median of 7 confirms this. According to the standard deviation participants were rather homogenous ($SD=2.32$); however the interquartile range suggests more diversity ($Q=4$). The minimum score reached was 4, while the maximum score was 12. Table 7.39 summarizes the descriptive information about the piloting data.

Value (s)	Total (12)
<i>M</i>	6.87
<i>MD</i>	7
<i>SD</i>	2.32
<i>Q</i>	4.0
<i>α</i>	0.63

Table 7.39: Descriptive results headword scoring Multiple Meaning Test (N=23)

Regarding facility values the Multiple Meaning Recognition Test of Academic Verbs achieved a satisfactory result. As Table 7.40 shows 66.67% of all items had a good facility value, 33.34% were difficult and 25% were easy.

above 0.7	25%
0.3-0.7	66.67%
below 0.3	33.34%

Table 7.40: Item facility headword scoring Multiple Meaning Test (N=23)

In a next step, distractors are investigated. Out of 24 distractors 14 worked well, being chosen by over ten percent of the participants. However, ten distractors worked rather

poorly, with a percentage below ten percent. The distractors were replaced in the final version that was used in the second data collection. Moreover, it was interesting to see that for the item *approach*, which had a very poor facility value of 0.96, both distractors did not work. The same accounts for *commit* and *promote*. For *assume*, *affect*, *perceive* and *attribute* only one distractor was chosen too infrequently.

Concerning item discrimination, the results are not as good as for facility value. As Table 7.41 shows items are evenly distributed to the categories with 25% each. Overall this means that 75% of all items pass the mark of a satisfactory discrimination value.

	Number of items	Percentage
<i>very good</i>	3	25.0%
<i>Good</i>	3	25.0%
<i>Marginal</i>	3	25.0%
<i>Poor</i>	3	25.0%

Table 7.41: Discrimination values headword scoring Multiple Meaning Test piloting

Moreover, it is interesting to see how clearly the facility values of the individual items correspond to the discrimination value and the reliability score (see Table 7.42). The potential link between these formulas has already been explained in section 7.2.

Item-Total Statistics						Facility values
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted	
<i>approach</i>	5.9130	5.265	0.087	0.714	0.634	0.957
establish	6.5652	4.257	0.470	0.463	0.570	0.304
maintain	6.3913	4.067	0.516	0.683	0.556	0.478
pursue	6.5652	4.530	0.320	0.684	0.602	0.304
impose	6.6522	4.601	0.339	0.540	0.599	0.217
assume	6.4348	4.621	0.236	0.545	0.620	0.435
affect	6.2609	4.202	0.460	0.632	0.570	0.609
perceive	6.2609	4.565	0.271	0.316	0.612	0.609
<i>commit</i>	5.9130	5.810	-0.460	0.300	0.678	0.957
attribute	6.2609	4.656	0.226	0.409	0.622	0.609
promote	5.9130	5.174	0.183	0.390	0.626	0.957
conceive	6.4348	4.439	0.326	0.377	0.600	0.435

Table 7.42: Relationship facility value, discrimination value and Cronbach alpha

As can be seen in Table 7.42 *approach* and *commit* are very easy for participants, have a poor discrimination value and would increase the reliability score if deleted. Moreover, for these two items both distractors did not work well. Consequently, it could be argued that all

arguments seem to indicate an exclusion of the items. Nonetheless, it has been decided to keep the items, since it is psychologically beneficial to put easy items at the beginning of the test to reassure the participants (Henning, 1987, p. 50).

The reliability value of the Multiple Meaning Recognition Test of Academic Verbs is $\alpha=0.63$ in the piloting phase. It can be argued that for such a small item number this is a satisfactory Cronbach alpha (Field, 2018). However, it will be seen, if adaptations made to the distractors of the test will change the reliability value of the Multiple Meaning Recognition Test in the second data collection.

Partial credit scoring

Similar to the headword scoring the data was not normally distributed for the partial credit scoring, [$D(23)=0.964$, $p=0.541$]. However, for partial credit the data was clearly normally distributed, which is the reason why only the mean will be used as reference point here. Out of a maximum score of 24 points participants reached on average 18.48 points, which equates 77%. Hence, it seems as if most participants were able to identify at least one correct meaning of the target words. This is confirmed by the minimum score of 14 and the maximum score of 24. The standard deviation of 2.73 reveals that for partial credit scoring the group was rather homogenous.

Value (s)	Total (24)
<i>M</i>	18.48
<i>SD</i>	2.73
α	0.67

Table 7.43: Descriptive statistics partial credit Multiple Meaning Test (N=23)

Regarding facility values the division in incorrect, partially correct answer and correct has been made.

incorrect	one correct answer	Correct
0.89%	40,22%	58.90%

Table 7.44: Item facility partial credit Multiple Meaning Recognition Test (N=23)

As visible in Table 7.44 only 0.89% of all answers were incorrect, meaning that both correct options were not chosen. Among the other two categories there is a near equal distribution, with 40.22% having one correct answer and one distractor and 58.90% with both correct answers. Hence, facility values appear to be comparable to the headword scoring method.

On the contrary, discrimination values are slightly differently distributed (see Table 7.45). Instead of an equal distribution, the partial credit method has fewer marginal and very good items but more good items.

	Number of items	Percentage
<i>very good</i>	2	16.67%
<i>Good</i>	6	50.0%
<i>Marginal</i>	1	8.33%
<i>Poor</i>	3	25.0%

Table 7.45: Discrimination values partial credit Multiple Meaning Recognition Test (N=23)

However, overall 75% achieve an at least marginal discrimination and only 25% a poor discrimination, which corresponds to the findings of the headword scoring.

Similar to facility values and item discrimination reliability is comparable to the headword scoring with a Cronbach $\alpha=0.67$. Consequently, it was decided that for the main data collection points partial credit scoring was used, since it had a slightly better reliability score and a more balanced discrimination.

7.4.4 2nd data collection⁵⁸

The Multiple Meaning Recognition Test of Academic Verbs was administered at the second data collection together with the Form-meaning Recall Test of Academic Verbs, the Form-meaning Recognition Test of Academic Verbs and the Collocation Recognition Test of Academic Verbs. Therefore, the information about participants and test administration procedure can be found in section 7.1.5.

Research questions

As for the three other tests of the test battery the same research questions as in the piloting were investigated one more time for the data of the 2nd data collection:

1. Is the Multiple Meaning Recognition Test of Academic Verbs satisfactory concerning difficulty?
2. Does the Multiple Meaning Recognition Test of Academic Verbs discriminate well enough between successful and less successful students?
3. Do the distractors of the Multiple Meaning Recognition Test of Academic Verbs work well enough?
4. Is the Multiple Meaning Recognition Test of Academic Verbs reliable?

⁵⁸ The Multiple Meaning Recognition Tests of Academic Verbs used for the 2nd and 3rd data collection can be found in appendix 7.1.17 and 7.1.18 together with their solutions.

Results⁵⁹

To answer the first research question measures of central tendency as of variability are calculated (see Table 7.46). Since the scores of the Multiple Meaning Recognition Test are not normally distributed, [$D(202)=0.938, p<0.001$], in the second data collection the median and the interquartile range are interpreted. However, the mean and the standard deviation are additionally reported for reasons of completeness. The comparison reveals that the mean score with 18.63 is very close to the median score of 19 points out of a maximum of 24 points reached. Therefore, the group appeared to be relatively proficient at average on the Multiple Meaning Recognition Test and the results of the second data collection align with the piloting results. Moreover, the group appears to be more or less homogeneous with an interquartile range of 3.00 and a standard deviation of 2.57.

Value (s)	Total (24)
<i>MD</i>	19.00
<i>M</i>	18.63
<i>Q</i>	3.00
<i>SD</i>	2.57
<i>α</i>	0.59

Table 7.46: Descriptive results for the Multiple Meaning Recognition Test, second data collection (N=204)

Since two outliers could be identified in the data set, the descriptive analysis has been repeated without the outliers but results show that there is no big difference. While the median stays at 19 points the mean score increases to 18.74. Similarly, the interquartile range stays at 3.00, while the standard deviation decreases slightly to 2.33. The distribution is still not normally distributed, [$D(200)=0.969, p<0.001$].

In addition to calculating the descriptive statistics presented above, research question 1 concerning the difficulty of the Multiple Meaning Recognition Test of Academic Verbs was attempted to be answered by the examination of facility values. Since the calculated differences for the group without outliers are too small to make any visible difference, only the numbers of the whole group are presented in the following.

As visible in Table 7.47 seven out of twelve items had a facility value between 0.3 and 0.7, which means that they can be considered suitable items. Only two items were difficult, and three items can be classified as easy.

⁵⁹ More detailed insights into the SPSS output of the analysis can be found in appendix 7.6.2.

above 0.7	3	25.00%
0.3-0.7	7	58.34%
below 0.3	2	16.67%

Table 7.47: Item facility Multiple Meaning Recognition Test, 2nd data collection (N=202)

Hence, facility values can be considered suitable, since some easy items should be placed at the beginning for psychological reasons. Moreover, since the same test is administered more than once more difficult items might be suitable for the next round of data collection.

While item facility appears to be suitable for the purpose of the study, discrimination values are not as satisfactory. As summarized in Table 7.48 only 66.68% of all items are satisfactory. There are no items with a very good discrimination. 33.34% of all items appear to discriminate rather poorly.

	Number of items	Percentage
Very good	0	0%
Good	4	33.34%
Marginal	4	33.34%
Poor	4	33.34%

Table 7.48: Discrimination values Multiple Meaning Recognition Test, 2nd piloting (N=202)

However, this number can partially be explained by two items, namely *approach* and *maintain*, which could be considered to be deleted from the test or exchanged for different items in the future.⁶⁰

8 Results on vocabulary tests

The following chapter will first explore the data distribution of the test scores through presenting the descriptive statistics of all three data collection points. After presenting the basic information for each of the points of investigation, it will be explored whether participants show a significant development of vocabulary knowledge over the three points in time for each of the word knowledge components. This corresponds to the following research questions:

- 🕒 Which development of academic verb knowledge is observable over the course of 1 academic year (T1-T3)?
 - Which development of form-meaning recognition knowledge of academic verbs is observable over the course of one academic year (T1-T3)?

⁶⁰ In general the Multiple Meaning Recognition Test of Academic Verbs should be extended in future studies, including more items, since this would enhance the probability of higher reliability.

- Which development of form-meaning recall knowledge of academic verbs is observable over the course of one academic year (T1-T3)?
- Which development of collocation recognition knowledge of academic verbs is observable over the course of one academic semester (T1-T3)?
- Which development of multiple meaning recognition knowledge of academic verbs is observable over the course of one academic semester (T1-T3)?

Similarly, the interrelatedness of the four word knowledge components for all three data collection points investigated will be analyzed, answering the following research questions:

 *Which relationship can be found for the four word knowledge components observed?*

 *How stable are relationships between the four word knowledge components observed with the help of correlational analysis and a SEM model?*

Therefore, the data gathered will also be utilized to examine the general nature of vocabulary knowledge by using structural equation modeling to examine whether a model inspired by Nation's (2013) taxonomy is suitable for the data set collected and how stable the model is over time.

8.1 1st data collection (T1)

Since the first data collection was conceived as being the baseline for further vocabulary development only the Form-meaning Recognition and the Form-meaning Recall Test of Academic Verbs, which are classically associated with the concept of vocabulary size, were distributed. For both tests a maximal score of 72 points was possible. As visible in Table 8.1 below this score was only reached for the Form-meaning Recognition Test of Academic Verbs, while the maximal score reached for the Form-meaning Recall Test of Academic Verbs was 60 points.⁶¹

Variable	N	Min.	Max.	M	Mdn	95% CI	SD	IQR
Form-meaning recognition	147	32.00	72.00	59.12	60.00	57.86; 60.37	7.70	11.00
Form-meaning recall	147	13.00	60.00	35.34	35.00	33.76; 36.92	9.69	14.00

Table 8.1: Descriptive statistics_1st data collection

As implied by the mean and median scores⁶² of both tests, the data distribution of the two word knowledge components is quite different. While data of the Form-meaning Recall Test is normally distributed, as shown by a Kolmogorov-Smirnov test ($D(147)=0.057$, $p=0.200$)

⁶¹ More detailed statistical output can be found in appendix 8.1.

⁶² The mean and the median are both referred to as “measures of central tendency”, which “describe the data set with a single numerical value” (Dörnyei, 2007, p. 213). While the mean is “the average of the scores”, the median is “the middle point in a set of scores that have been arranged in a rank order” (Dörnyei, 2007, p. 214). While the mean is sensitive to extreme scores, the median is not sensitive to extreme scores, which is why it is reported additional to the mean if no normal distribution is given (Dörnyei, 2007).

and a Shapiro-Wilk test ($W(147)=0.993, p=0.653$), depicted by a nice bell-shaped curved in the histogram below, the Form-meaning Recognition Test is subdued to a ceiling effect, impeding a normal distribution ($D(147)=0.099, p=0.001$) as visible in Figure 8.1 below.

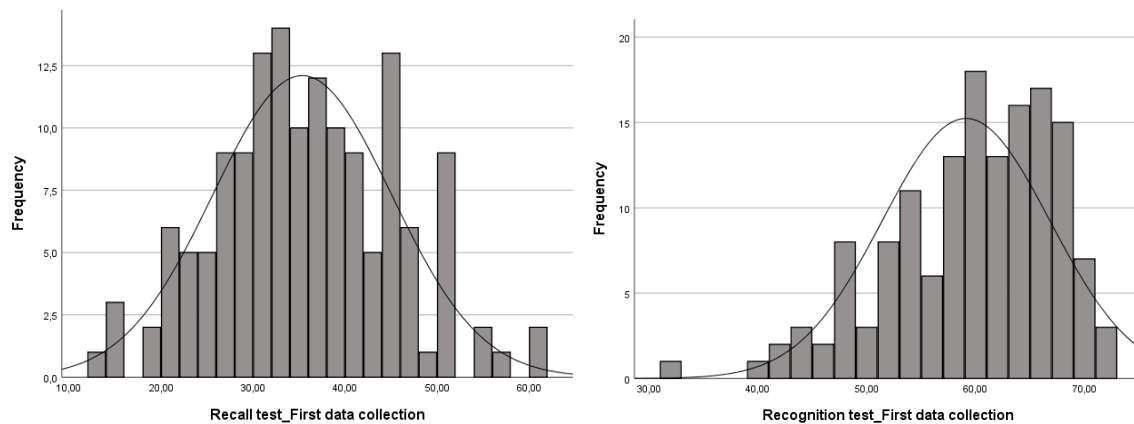


Figure 8.1: Data distribution of Form-meaning Recall and Recognition Test_1st data collection

Conversely to the overall data distribution, measures of variance (SD & IQR) are comparably large, indicating a heterogeneous answer behavior within the received distribution.

Additional to the descriptive properties of the tests' interrelatedness of the word knowledge components was investigated by carrying out a Spearman's correlation⁶³ due to the non-normal distribution of the form-meaning recognition data (Field, 2018). As visible in Table 8.2 below the receptive and productive form-meaning word knowledge components are strongly related (Field, 2018, p. 340).

	Form-meaning recognition	Form-meaning recall
Form-meaning recognition	-	0.73**
Form-meaning recall	0.73**	-

**Correlation is significant at the 0.01 level

Table 8.2: Correlations between form-meaning recognition and recall_1st data collection

Hence, it can be argued that the scores on the Form-meaning Recognition Test of Academic Verbs and the Form-meaning Recall Test of Academic Verbs show a high degree of relationship, which implies that these two word knowledge components are highly interconnected.⁶⁴

⁶³ Spearman's correlation or more precisely Spearman's rank order correlation is a non-parametric correlation, which "is based on the ranks of the data (in a rank order) rather than the actual values" (Dörnyei, 2007, p. 230). This makes it suitable for ordinal data or interval data, which is not normally distributed.

⁶⁴ It has to be considered that a correlation cannot identify cause and effect (Dörnyei, 2007, p. 225), which means that results only give information about interrelatedness.

8.2 2nd data collection (T2)

Additional to the two tests of form-meaning knowledge, two other tests were applied in the second data collection, namely the Multiple Meaning Recognition Test of Academic Verbs and the Collocation Recognition Test of Academic Verbs. While the Collocation Recognition Test has a maximal score of 26 points, participants can only reach a maximum of 24 points in the Multiple Meaning Recognition Test. Similar to the Form-meaning Recognition Test, which keeps its maximal score of 72, at least one participant reached the maximum score available in the Multiple Meaning Recognition Test.⁶⁵

Variable	N	Min.	Max.	M	Mdn	95%CI	SD	IQR
Form-meaning recognition	147	30.00	72.00	59.63	62.00	58.30; 60.95	8.12	11.00
Form-meaning recall	147	16.00	65.00	40.76	41.00	39.12; 42.41	10.09	14.00
Multiple meaning recognition	147	13.00	24.00	18.72	19.00	18.35; 19.09	2.27	3.00
Collocation recognition	147	8.00	24.00	16.34	17.00	15.75; 16.93	3.64	6.00

Table 8.3: Descriptive statistics_2nd data collection

For the Form-meaning Recall Test a descriptive increase of 5 points is visible in comparison to the first data collection, with a maximum score of 65 points reached.

While the data distribution of the two Form-meaning tests is comparable to the first data collection, with the Form-meaning Recall Test of Academic Verbs showing a normal distribution ($D(147)=0.059$, $p=0.200$) in the Kolmogorov-Smirnov test⁶⁶ and the Form-meaning Recognition Test of Academic Verbs being non-normal ($D(147)=0.118$, $p<0.001$), both other tests are also non-normally distributed ($D(147)=0.137$, $p<0.001$; $D(147)=0.098$, $p=0.002$). However, no striking ceiling effect is visible in comparison to the Form-meaning Recognition Test as visible in Figure 8.2.

⁶⁵ More detailed statistical output can be found in appendix 8.2.

⁶⁶ While the use of the Kolmogorov-Smirnov test and the Shapiro-Wilk test to reach a decision about normal distribution is widely spread, Field (2018, p. 250) advises to additionally plot the existent data to reach an informed decision about normal distribution and its extend. Consequently, skewness and kurtosis of a data distribution curve should be interpreted as well.

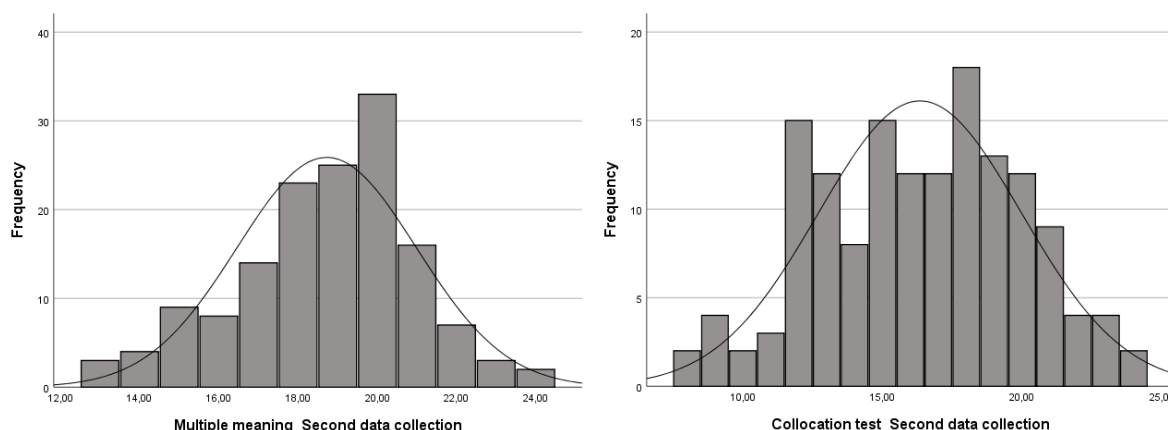


Figure 8.2: Data distribution of Multiple Meaning and Collocation Recognition Test_2nd data collection

Moreover, while the variance appears to be relatively low in comparison to the Form-meaning link tests at first sight, group answers are still relatively heterogeneous, since one has to contemplate these measures in proportion to the maximum score. Hence, since the maximum score is much lower for the Multiple Meaning Recognition and the Collocation Recognition Test, with 24 and 26 points, than for the Form-meaning link tests with 72 points, the variance is rather large within the groups.

Taking a look at the interrelatedness of the four word knowledge components measured reveals that while the correlation coefficient between form-meaning recognition and form-meaning recall has decreased from the first to the second data collection, both tests still seem to be highly correlated with each other (see Table 8.4).

	(1)	(2)	(3)	(4)
Form-meaning recognition (1)	-			
Form-meaning recall (2)	0.68**	-		
Multiple Meaning recognition (3)	0.52**	0.47**	-	
Collocation recognition (4)	0.65**	0.50**	0.46**	-

**Correlation is significant at the 0.01 level

Table 8.4: Correlations between word knowledge components_2nd data collection

The same accounts for the Form-meaning Recognition Test and results on the Collocation Recognition Test, which also appear to be highly related. In comparison the other word knowledge components are not as strongly correlated with each other but still display medium correlation scores (Field, 2018, p. 340). Hence, it can be said that the word knowledge components under investigation appear to be interrelated; however to varying degrees.

8.3 3rd data collection (T3)

Comparing the average scores of the second data collection to the values of the third data collection in Table 8.5 reveals that descriptively only two tests depict a positive development regarding the average score achieved.⁶⁷

Variable	N	Min.	Max.	M	Mdn	95% CI	SD	IQR
Form-meaning recognition	147	39.00	72.00	61.95	63.00	60.71; 63.18	7.55	9.00
Form-meaning recall	147	22.00	67.00	44.04	44.00	42.89; 45.92	9.30	14.00
Multiple meaning recognition	147	12.00	24.00	19.30	19.00	18.92; 19.68	2.35	3.00
Collocation recognition	147	6.00	25.00	17.16	17.00	16.55; 17.77	3.74	5.00

Table 8.5: Descriptive statistics_3rd data collection

While both the Form-meaning Recognition and Recall Test have increased from 59.63 to 61.95 points and respectively 41 to 44 points, the Multiple Meaning Recognition Test and the Collocation Recognition Test have stagnated at an average of 19 and 17 points achieved. This corresponds to a drop of variance within the Form-meaning tests from an *IQR* of 11 to 9 points and an *SD* of at least 10.09 to 9.30 points.

Concerning the overall data distribution, no changes can be detected, with only the Form-meaning Recall Test being normally distributed according to the Kolmogorov-Smirnov test ($D(147)=0.055, p=0.200$).

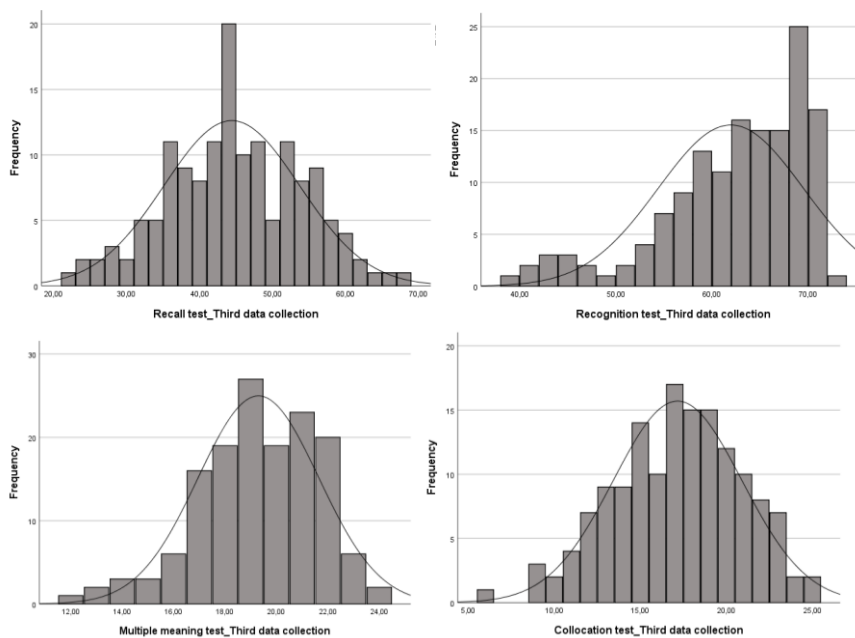


Figure 8.3: Data distribution of all four tests_3rd data collection

⁶⁷ The statistical output of the third data collection can be found in appendix 8.3.

Nevertheless, similar to the second data collection only the Form-meaning Recognition Test depicts an even enhanced ceiling effect, while the other two tests appear to be fairly near to a normal distribution as visible in Figure 8.3 above.

Regarding correlational patterns the degree of interrelatedness of the individual word knowledge components appears to remain largely unchanged from the second to the third data collection as visible in Table 8.6 below.

	(1)	(2)	(3)	(4)
Form-meaning recognition (1)	-			
Form-meaning recall (2)	0.73**	-		
Multiple Meaning recognition (3)	0.54**	0.47**	-	
Collocation recognition (4)	0.62**	0.47**	0.46**	-

**Correlation is significant at the 0.01 level

Table 8.6: Correlations between word knowledge components_3rd data collection

8.4 Development of word knowledge components

As already foreshadowed in chapters 8.1-8.3 through descriptive changes in the average test scores reached, this chapter investigates whether participants display statistically significant developments regarding their knowledge of the four different word knowledge components from the 1st over the 2nd and to the 3rd data collection point. For the Form-meaning Recognition and Recall Tests of Academic Verbs, encompassing three data collection points, this is attempted by calculating Repeated Measures ANOVAs, which is “used when the same entities participate in all conditions of an experiment or provide data at multiple time points” (Field, 2018, p. 651). Similar to other inferential statistics Repeated measures ANOVAs require normally distributed interval scales. However, this assumption is only given for the Form-meaning Recall Test, which has been normally distributed over all three data collection points (see chapters 8.1-8.3). Hence, a robust Repeated measures ANOVA (Field, 2018, p. 676) was calculated in R with the package WRS2 (Mair, Wilcox & Patil, 2005) for the Form-meaning Recognition Test. In this package a heteroscedastic one-way repeated measures ANOVA with a trim factor of 0.2 is used to counteract non-normally distributed data (Mair & Wilcox, 2021, p. 24). This means that 10% of the data at the two extreme ends are cut to balance the distribution more, which appears to be a suitable way to react to the observed ceiling effect in the data of the Form-meaning Recognition Test. Additionally, development concerning the Multiple Meaning Recognition Test and the Collocation Recognition Test, was measured on basis of a Wilcoxon test, a non-parametric paired samples t-test, since these tests were only utilized at two data collection points (Field, 2018, p. 298). In the following the results will be presented for all four tests.

8.4.1 Form-meaning Recognition Test of Academic Verbs

Overall, a statistically significant development, with a small effect size ($\eta^2=0.21$)⁶⁸ of the form-meaning recognition knowledge over the timespan of one year could be detected ($F(1.99, 174.19)=22.95, p<0.001$)⁶⁹. Moreover, as visible in Figure 8.4 below participants did not improve significantly from the first to the second data collection point ($p=0.050$) but instead between the second and third data collection ($p=0.025$) and overall between the first and the third data collection ($p=0.017$).

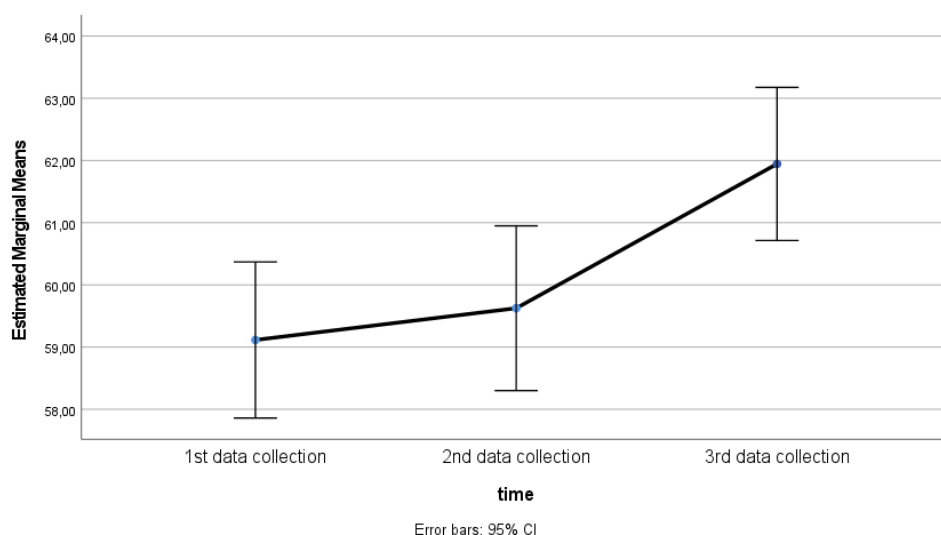


Figure 8.4: Development of form-meaning recognition knowledge

This initial phase of stagnation might however be due to the ceiling effect visible in the data already in the first data collection. Due to the already very advanced level of form-meaning recognition knowledge at the first data collection, it might have been more difficult to improve as quickly as for other word knowledge components.

Moreover, it becomes visible that the data sets include an increasing number of outliers the longer the study lasted (see Figure 8.5 below).

⁶⁸ η^2 = "Eta-squared is commonly used in ANOVA and t test designs as an index of the proportion of variance attributed to one or more effects. The statistic is useful in describing how variables are behaving within the researcher's sample. In addition, because eta-squared is a measure of effect size, researchers are able to compare effects of grouping variables or treatment conditions across related studies." (Fay & Boyd, 2010)

⁶⁹ For this calculation a Robust Repeated measures ANOVA was used. The R-code for the calculation can be found in appendix 8.4.1.

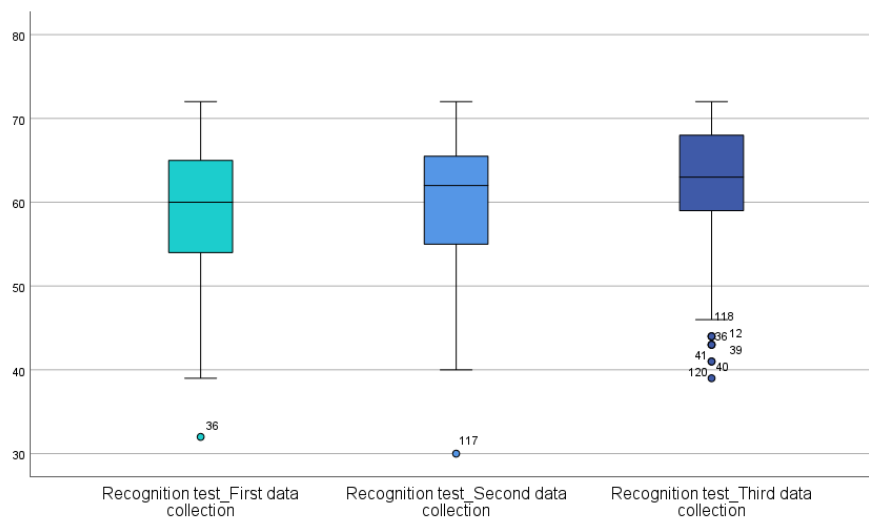


Figure 8.5: Development and outliers of Form-meaning Recognition Test of Academic Verbs

While it was decided not to exclude outliers in general, since the definition of an outlier is quite strict in a boxplot. Moreover, to value the authenticity of the data collected, the ascending number of outliers in the third data collection should also be interpreted in comparison to the other two data collection points in which they would not be counted as outliers yet. Overall a significant improvement concerning form-meaning recognition knowledge is visible.

8.4.2 Form-meaning Recall Test of Academic Verbs

Divergent to the Form-meaning Recognition Test, the Form-meaning Recall Test met all assumptions for a standard repeated measures ANOVA.⁷⁰ Firstly, as visible in Figure 8.6 below the Form-meaning Recall Test contains no outliers and is normally distributed.

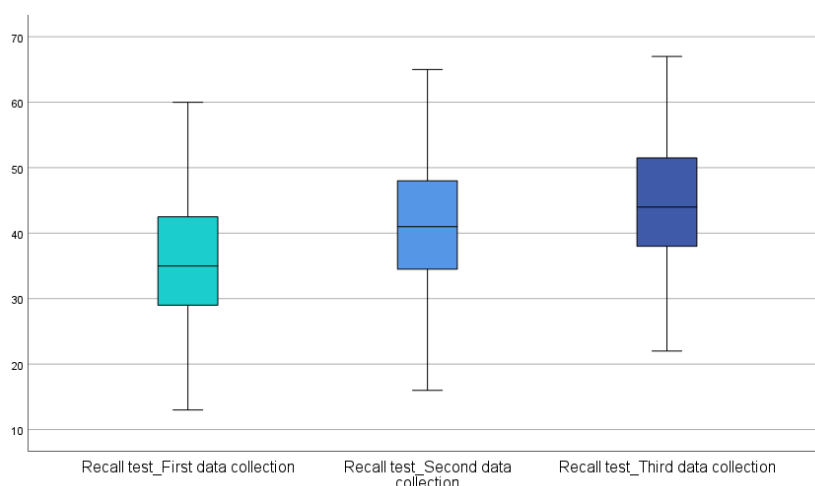


Figure 8.6: Development of Form-meaning Recall Test of Academic Verbs

Therefore, the assumption of sphericity needs to be checked. As Mauchly's sphericity test ($\chi^2(2)=14.69$, $p=0.001$) shows, sphericity is not given for the Form-meaning Recall Test data.

⁷⁰ See more detailed statistical output in appendix 8.4.2.

Consequently, either the Greenhouse-Geisser or the Huynh-Feldt correction for sphericity can be applied. Since the Greenhouse-Geisser effect is known to underestimate epsilon (ϵ) when the value is close to 1, it was decided to use the Huynh-Feldt correction for sphericity (Field, 2018, p. 655).

Considering these adjustments a statistically significant large effect ($\eta^2=0.50$) could be found for the development of form-meaning recall knowledge over the timespan of a year ($F(1.85, 269.53) = 147.88, p < 0.001$) as visible in Figure 8.7.

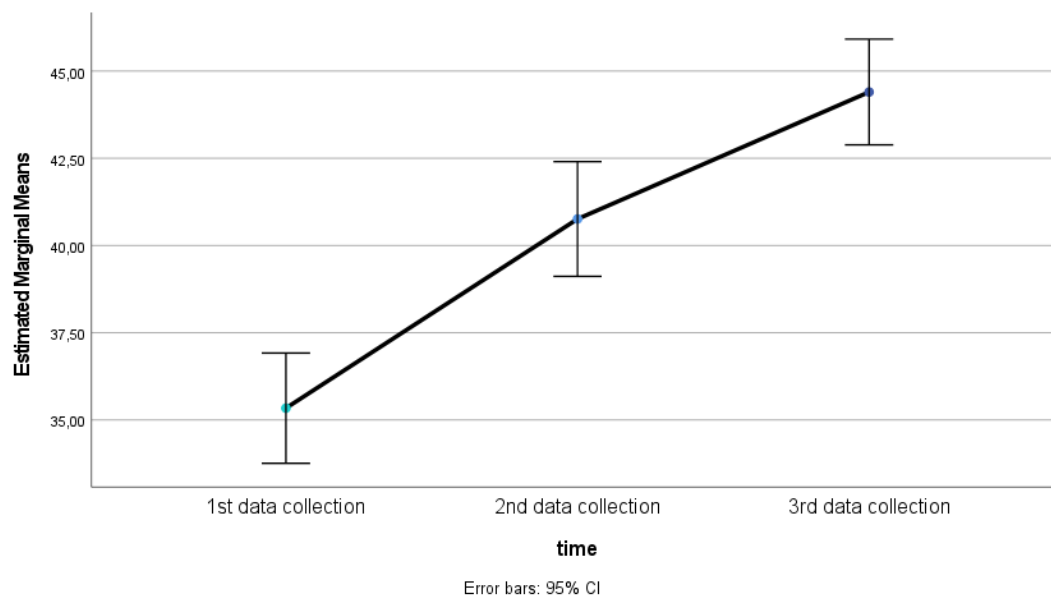


Figure 8.7: Development of form-meaning recall knowledge

Hence, 50.3% of the visible change in the data set over the three data collection points can be attributed to the factor of time. Moreover, a post hoc pairwise comparison using the Bonferroni correction shows that all pairwise comparisons are significant ($p < 0.001$).

8.4.3 Multiple Meaning Recognition Test of Academic Verbs

For the analysis of the Multiple Meaning Recognition Test the Wilcoxon test was used, which is a non-parametric related samples t-test comparing two data samples, since the assumption of normal distribution was not given despite the relatively large sample size of $N=147$.⁷¹

As visible in Figure 8.8 a significant increase of multiple meaning recognition knowledge can be detected between data collection two and data collection three ($Z=2.82, p=0.005$).

⁷¹ More detailed statistical output can be found in appendix 8.4.3.

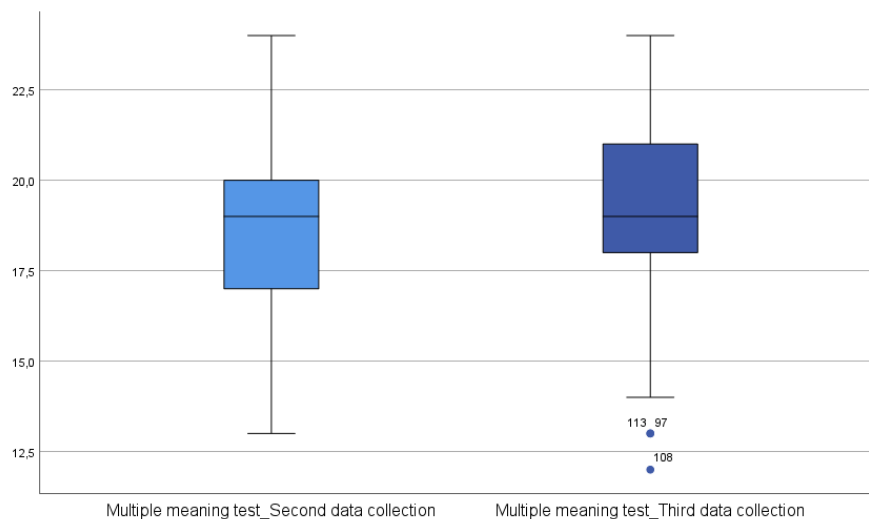


Figure 8.8: Development of multiple meaning recognition knowledge

However, the effect size, using the measure of r (Field, 2018, p. 319), is rather small with $r=0.23$, indicating a rather flat development of multiple meaning knowledge.

As for the Form-meaning Recognition Test of Academic Verbs, the Multiple Meaning Test of Academic Verbs displays an increase in outliers. While no outliers can be detected for the second data collection, three outliers can be identified for the third data collection, again indicating that some participants might have not been able to adapt to the pace of development of the rest of the group.

8.4.4 Collocation Recognition Test of Academic Verbs

Similar to the Multiple Meaning Recognition Test a Wilcoxon test is used to examine the development of collocation recognition knowledge.⁷² As Figure 8.9 indicates a statistical significant improvement of participants' receptive collocation knowledge can be found ($Z=3.90$, $p<0.001$).

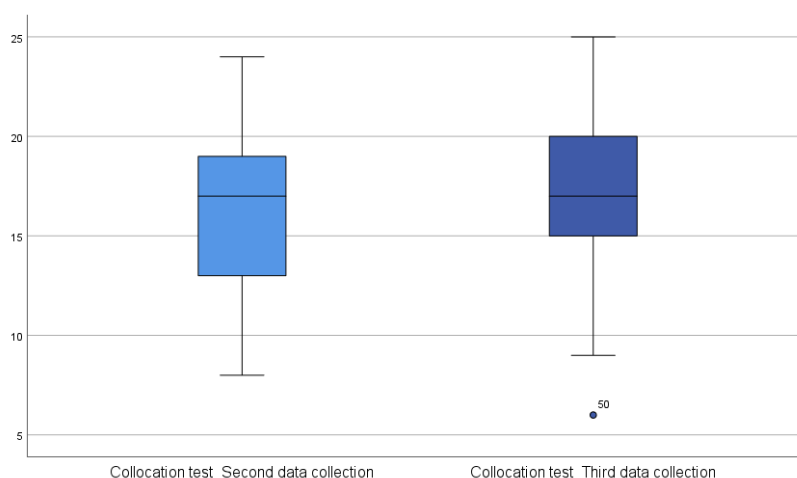


Figure 8.9: Development of collocation recognition knowledge

⁷² More detailed statistical output can be found in appendix 8.4.4.

Moreover, applying the same two measures as for the Multiple Meaning Recognition Test results in a medium effect size of $r = 0.32$. Hence, participants made a medium improvement in the area of collocation knowledge.

Overall, statistically significant improvements of all four word knowledge components could be detected over the time span of one academic year. However, only the development of the Form-meaning Recall Test showed a large effect and significance over all three data collection points, while the Form-meaning Recognition, the Multiple Meaning Recognition Test and the Collocation Recognition Test displayed a significant development but small to medium effect sizes.

8.5 SEM Models

While regarding each of the measured word knowledge components as individual area of vocabulary development was already quite fruitful, it might be even more insightful to investigate how these word knowledge aspects are connected to each other and how they relate to the overall construct of vocabulary knowledge. This was attempted by using the method of structural equation modeling (SEM).

Structural equation modeling (SEM) encompasses many statistical techniques, such as MANOVA, regression analysis or confirmatory factor analysis. In general SEM is used to model “relationships between observed variables [i.e. being variables measured in a study] and hypothesized underlying latent variables or both” (Schoonen, 2013, p. 1). Observed variables also termed measured, indicator or manifest variables are traditionally visually represented by squares or rectangles by researchers, while latent or unobserved variables are depicted as circles or ovals (Schreiber et al., 2006, p. 323). Amongst other aspects exactly the possibility to investigate not only the relationships between measured variables but also latent, theoretically driven variables, differentiates SEM from other statistical techniques such as path modeling (Schreiber et al., 2006, p. 323). Hence, SEM is a confirmatory statistical procedure (Sawaki, 2013, p. 1), such as an exploratory factor analysis, which researchers use to evaluate the fit of their proposed model/s regarding the fit on the existent, collected data (Nachtigall et al., 2003, p. 5; Schoonen, 2013).

This model fit is expressed by various different fit indices, which can be broadly differentiated in absolute and incremental or comparative fit indices (Hu & Bentler, 1999; von Eye, Spiel & Wagner, 2003) as can be seen in Table 8.7.

Absolute fit indices	Comparative or incremental fit indices
χ^2 -Test	CFI (comparative fit index)
χ^2/df	RMSEA (root mean squared error of approximation)
GFI (goodness-of fit)	TLI (Tucker-Lewis index)
AGFI (adjusted goodness-of fit)	
SRMR (standardized root mean square of residual)	

Table 8.7: Selection of fit indices for SEM Model

While absolute fit indices “directly assess how well an a priori model reproduces the sample data” (Hu & Bentler, 1999, p. 426), incremental fit indices compare the proposed model to a so-called baseline model. This baseline model most commonly assumes that all of the measured variables are allowed to have variances but that they are unrelated to each other (Hu & Bentler, 1999).

Starting with the most commonly used absolute fit index, namely the chisquare test (χ^2 -Test), this test expresses how large the difference between the observed co-variance matrix of the given data and the estimated population matrix of the hypothesized model is (Schoonen, 2013). Hence, one can remember as a rule of thumb that the smaller χ^2 the better the model fit. However, the χ^2 Test is influenced by sample size, number of variables included in the model and degrees of freedom (df) (Schoonen, 2013). To counteract this bias “the ratio of χ^2 and df [χ^2/df ,] is often used as a more descriptive measure” (Schoonen, 2013, p. 4; Ullman, 2007), which should not be larger than 2-3 (Schermelleh-Engel, Moosbrugger & Müller, 2003; Schreiber et al., 2006).

Another absolute fit index is the GFI or the AGFI, which can reach a maximum score of 1. However, these two estimators have been criticized for their sensitivity to model misspecification, small sample size, estimation method and violation of normality and independence (Hu & Bentler, 1999).

Finally, the SRMR is another frequently reported fit index for SEM models. However, the degree of strictness applied regarding this value has varied in the existent literature. While Schoonen (2013) and Schreiber et al. (2006) report that the SRMR needs to be at least smaller than 0.08, Hu and Bentler (1999, p. 449) have suggest an even stricter boundary at 0.05.

Considering the selection of comparative fit indices reported in Table 8.7 above, the CFI, ranging between values of 0-1, is frequently used for SEM evaluation. As implied by the category the CFI “indicates the relative improvement of the tested model over a baseline model, the so-called independence model” (Schoonen, 2013, p. 4), which assumes that all variables measured are uncorrelated and hence independent from each other. While in general a value of ≥ 0.95 is accepted as good model fit (Schoonen, 2013; Schreiber et al.,

2006), Schermelleh-Engel, Moosbrugger and Müller (2003) state that the cut off value should be lifted to ≥ 0.97 . Similar to the CFI the TLI is required to be relatively large with TLI ≥ 0.95 being a good model fit (Schreiber et al., 2006, p. 330).

While many fit indices are vulnerable to sample size, the RMSEA is robust to sample size and indicates the error of approximation. However, similar to other measures different cut-off values exist, with Schreiber et al. (2006) proposing the broadest and therefore most lenient boundary, with the RMSEA being between 0.06 and 0.08. Schoonen narrows this range to ≤ 0.06 and Hu and Bentler (1999) again suggesting the strictest boundary with RMSEA ≤ 0.05 .

As visible in the compendium of different fit indices above none of these measures should be used exclusively to evaluate a SEM model. Instead several fit indices suitable for the model and the underlying data should be reported to counteract the critique that each of the measures has received over the time. Hence, for this study the GFI and the AGFI will be neglected, since they are vulnerable to non-normal distribution of variables, which is given for most of our observed variables in the models. However, all other fit indices introduced above will be reported.

Similar to most other complex statistical measures, SEM models require the testing of assumptions and consequent suitable decisions on measurement methods. To apply the most common type of estimating parameters, namely the Maximum Likelihood Method (ML), adequate sample size as well as multivariate normally distributed interval variables are required (Nachtigall et al., 2003, p. 7). However, it is possible to perform SEM also on non-normal and/or ordinal variables with the help of adjusted fit indices, distribution free estimation, and/or bootstrapping (Schoonen, 2013; van Eye, Spiel & Wagner, 2003). This includes methods, such as the Weighted Least squares (WLS) or maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic (MLM). Since WLS requires very large sample size (Nachtigall et al., 2003, p. 7), which is not available in the present study, MLM has been chosen as estimator method for this research.

While being a very flexible and rich statistical approach, SEM has not been used extensively in the field of Applied Linguistics until now (Schoonen, 2013, p. 1). Hence, this study is one of the first studies in this field working with a SEM modeling approach. As visible in Figure 8.10 below two competing models of vocabulary knowledge are proposed:

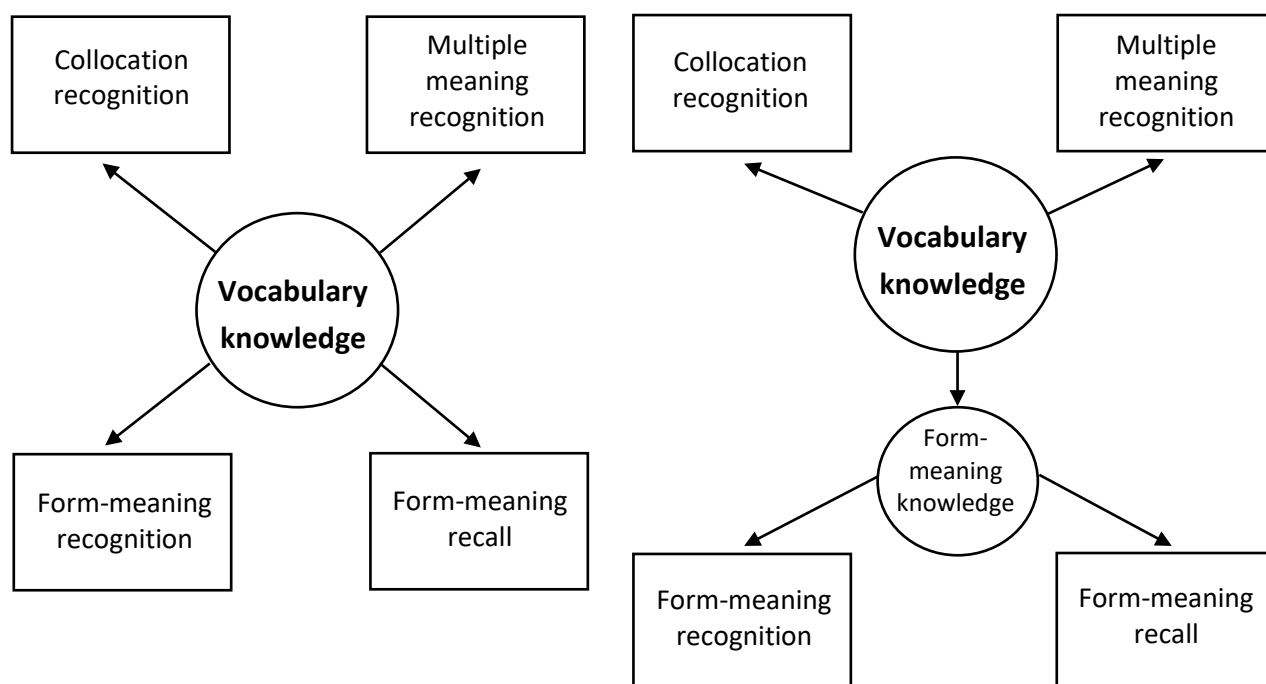


Figure 8.10: Two competing models of vocabulary knowledge

While the first model assumes that all four word knowledge components measured are equally related to the concept of vocabulary knowledge, Model 2 is aligned to Nation's (2013) taxonomy as explained in chapter 2.4, which proposes that form-meaning recognition and form-meaning recall together make up the category of form-meaning knowledge. Hence, it is indirectly assumed that the receptive and productive side of one word knowledge component are more closely connected than the individual word knowledge components to each other by proposing that each word knowledge component consists of a receptive and a productive side.⁷³

As already suggested in the descriptive statistics in chapters 8.1-8.3 the assumption of univariate normality was violated for all variables but form-meaning recall, calling for maximum likelihood parameter estimates with standard errors and a mean-adjusted chi-square test statistic (e.g. Satorra-Bentler corrections), which is considered to be robust against a non-normal distribution of variables. Regarding factorability the Kaiser-Meyer-Olkin (KMO) test suggests that a moderate factorability ($KMO > 0.70$; Field, 2018) is given ($KMO = 0.78$).

After checking for all assumptions, the SEM analyses was run for both types of models with the data of the second and the third data collection. Hence, it will be seen whether model fit indices remain stable over half a year between the two data collection points and whether the same model is suitable at both points in time. Therefore, assumptions about the stability and consistency of model fit can be made as well. Table 8.8 below depicts the fit indices for the two proposed models for the second data collection.

⁷³ The R codes for the SEM models can be found in appendix 8.5.

<i>Model</i>	<i>X²</i>	<i>df</i>	<i>p value</i> <i>p>0.05</i>	<i>X²/df</i> <i><3</i>	<i>CFI</i> <i>>0.97</i>	<i>TLI</i> <i>>0.95</i>	<i>RMSEA, 90% CI</i> <i><0.05/<0.08</i>	<i>SRMR</i> <i><0.05</i>
Model 1	2.385	2	0.304	1.19	0.998	0.994	0.036, <i>p</i> =0.428	0.022
Model 2	0	1	0.996	0	1.0	1.033	0.000, <i>p</i> =0.999	0

Table 8.8: Model-fit indices for SEM Model 1 & 2_Second data collection

As visible Model 2 appears to have the better model fit based on all fit indices considered for the analysis for the data collected during the second data collection. However, considering the relatively small sample size for such an analysis (N=147) the interpretation mainly relies on the standardized root mean square (SRMR) and the comparative fit index (CFI) (Hu & Bentler, 1999).

Regarding the data from the third data collection the results for both proposed models can be found in Table 8.9.

<i>Model</i>	<i>X²</i>	<i>df</i>	<i>p value</i> <i>p>0.05</i>	<i>X²/df</i> <i><3</i>	<i>CFI</i> <i>>0.97</i>	<i>TLI</i> <i>>0.95</i>	<i>RMSEA, 90% CI</i> <i><0.05/<0.08</i>	<i>SRMR</i> <i><0.05</i>
Model 1	3.365	2	0.186	1.683	0.994	0.989	0.063, <i>p</i> =0.293	0.024
Model 2	0.988	1	0.320	0.988	1.0	1.0	0.000, <i>p</i> =0.412	0.012

Table 8.9: Model-fit indices for SEM Model 1 & 2_Third data collection

Again Model 2 appears to be more suitable at first sight with the result of the X^2 Test being lower, indicating a smaller gap between the proposed model and the existent data. Moreover, Model 2 has the largest CFI with the maximum value of 1.0 and a smaller SRMR with 0.012.

In a next step, the standardized regression coefficients (β values) for all regression paths need to be considered before finally deciding on a model. Table 8.10 reports all standardized regression coefficients for Model 1 for the second and the third data collection.

Latent variable	Direction	Observed variable	Second data collection		Third data collection	
			B	p	β	p
Vocabulary knowledge	→	Collocation recognition (COL_Recog)	0.708	<0.001	0.639	<0.001
Vocabulary knowledge	→	Form-meaning recognition (FM_Recog)	0.875	<0.001	0.892	<0.001
Vocabulary knowledge	→	Form-meaning recall (FM_Recall)	0.741	<0.001	0.795	<0.001
Vocabulary knowledge	→	Multiple meaning recognition (MM_Recog)	0.594	<0.001	0.553	<0.001

Table 8.10: Summary of standardized regression coefficients for Model 1 for the second and third data collection

As visible all regression paths were found to be significant with the strongest connection found between Vocabulary knowledge and Form-meaning recognition for both points in time. Hence, the latent variable of Vocabulary knowledge appears to be strongly connected to knowledge of Form-meaning recognition of academic verbs. Moreover, the regression path between the latent variable Vocabulary knowledge and Form-meaning recall was the second strongest for both data collections. However, the distance between the other two word knowledge components, Collocation recognition and Multiple meaning recognition knowledge, has increased over time, since the regression coefficients for the observed variable Collocation recognition decreased from the second to the third data collection. The least connection was found between Vocabulary knowledge and Multiple meaning recognition for both points in time.

As visible in Figure 8.11, which depicts the SEM of Model 1, the total variance, which remains unexplained by the latent variable of Vocabulary knowledge (visualized by curved arrows), is related to the strength of the regression coefficients.

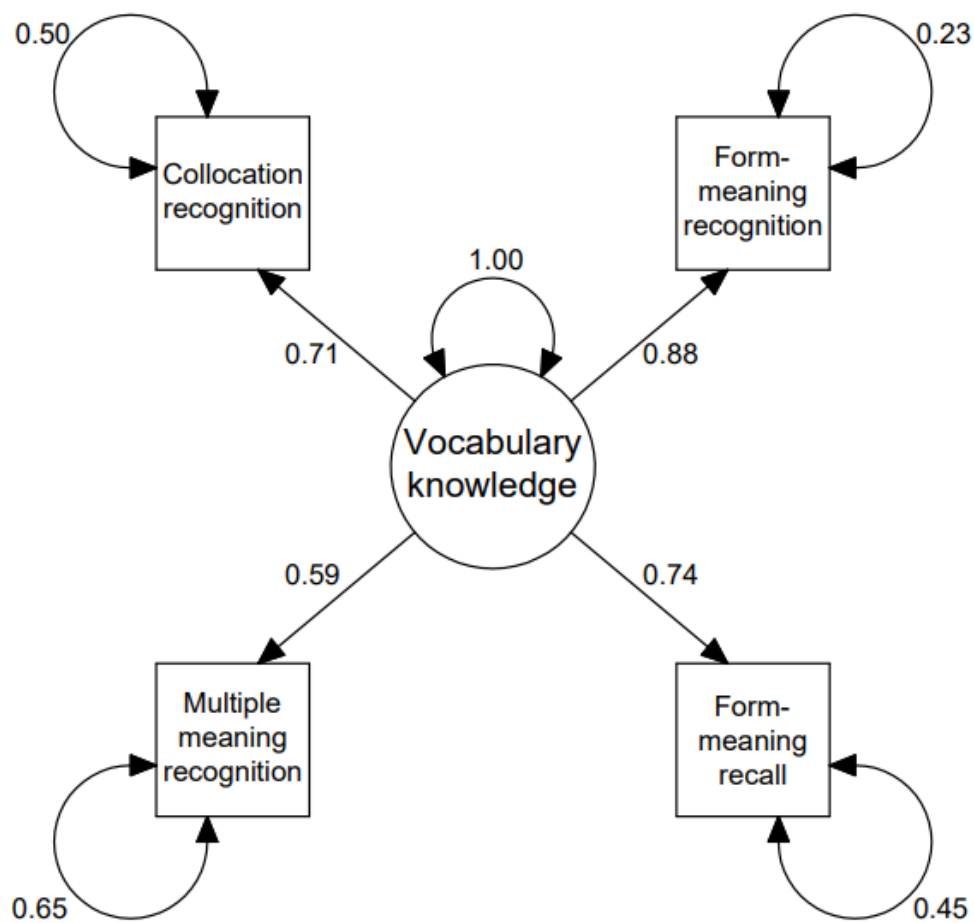


Figure 8.11: Model 1 of Vocabulary knowledge_Second data collection

Hence, the observed variable of Form-meaning recognition with strong β scores has the smallest unexplained total variance with 0.23 for the second data collection and 0.20 for the third data collection. Conversely, Multiple meaning recognition with the smallest β scores has the highest amount of unexplained total variance, namely 0.65 and 0.69. The same phenomenon can also be found for Model 2 as visible in Figure 8.12 depicting the SEM of Model 2.

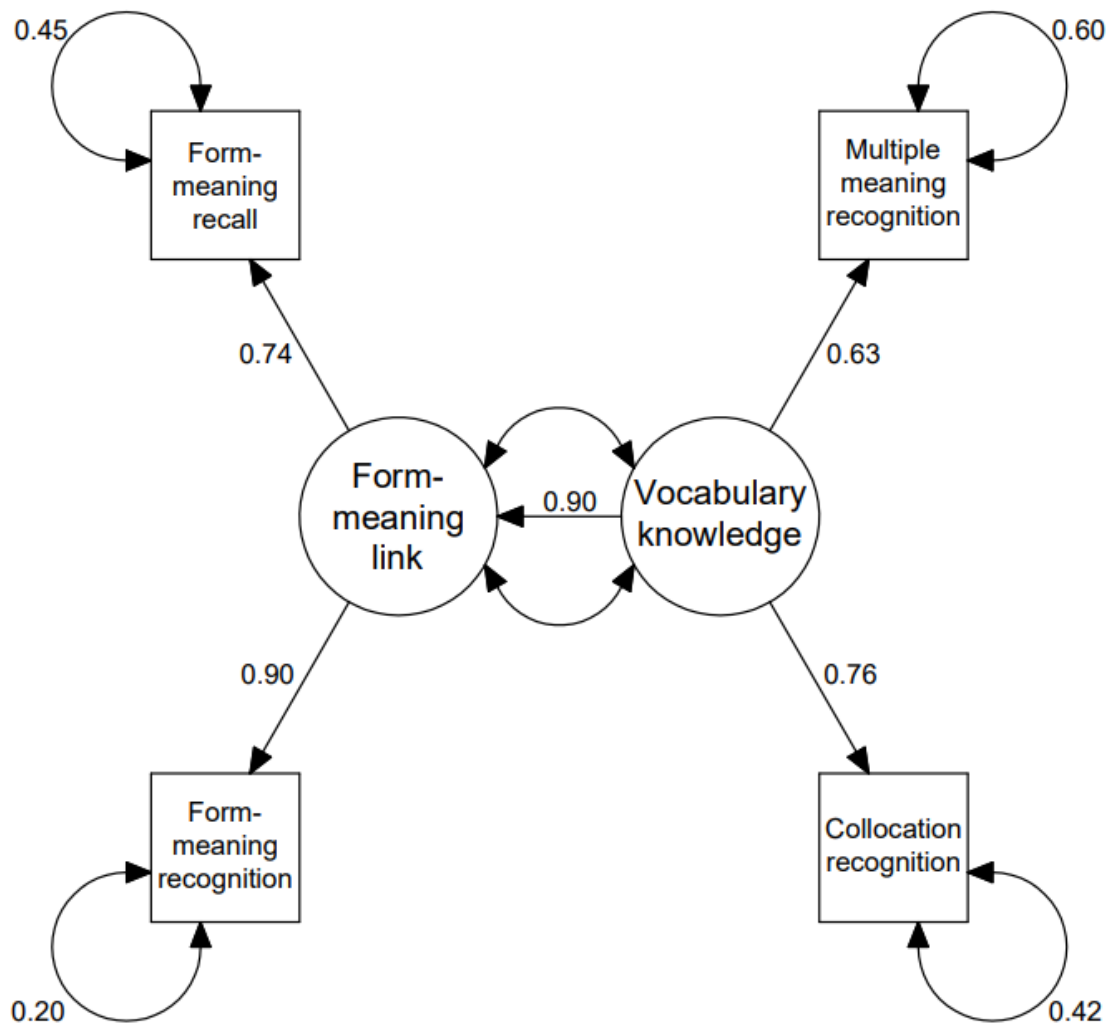


Figure 8.12: Model 2 of Vocabulary knowledge_Second data collection

Similar as for Model 1 Table 8.11 summarizes the standardized regression coefficients (β values) for all regression paths for both points in time for Model 2.

Latent variable	Direction	Latent variable	Second data collection (T1)		Third data collection (T2)	
			β	P	β	P
Vocabulary knowledge	→	Form-meaning knowledge (FM)	0.900	<0.001	0.897	<0.001

Latent variable	Direction	Observed variable	Second data collection (T1)		Third data collection (T2)	
			β	P	β	P
Vocabulary knowledge	→	Collocation recognition (COL_Recog)	0.762	<0.001	0.698	<0.001
Vocabulary knowledge	→	Form-meaning recognition (FM_Recog)	0.895	<0.001	0.904	<0.001
Vocabulary knowledge	→	Form-meaning recall (FM_Recall)	0.738	<0.001	0.791	<0.001
Vocabulary knowledge	→	Multiple meaning recognition (MM_Recog)	0.630	<0.001	0.596	<0.001

Table 8.11: Summary of standardized regression coefficients for Model 2 for the second and third data collection

As visible there is a very strong regression path between the two latent variables Vocabulary knowledge and Form-meaning knowledge (FM), which indicates that the two latent variables can be considered to be nearly identical. Hence, while Model 2 appears to have the better model fit on the data, the regression coefficient between the two latent variables is too high to be regarded as two independent constructs.

Therefore, the overarching category of Form-meaning knowledge for the two observed variables Form-meaning recognition and Form-meaning recall becomes obsolete and leaves us with Model 1 as the most suitable model for the data collected.

While this does not allow any conclusions about the comparability of relationships between recognition and recall knowledge of individual word knowledge components and the relationships between different word knowledge components, this result shows that most of the participants' vocabulary knowledge of academic verbs appears to be connected to Form-meaning knowledge, with Form-meaning recognition having an even stronger influence than Form-meaning recall. In contrast, word knowledge components, which are more associated with the vocabulary depth appear to be explained less by this construct of

vocabulary knowledge. Hence, it remains to be seen which factors are able to explain the remaining total variance not explained by the construct of vocabulary knowledge for these word knowledge components.

Concerning the aspect of time, model fit appears to remain fairly stable over the two data collection points, with regression coefficients of the individual paths changing slightly but being stable regarding the degree of strength detectable. Hence, it can be suggested that Model 1 is a very suitable model for vocabulary knowledge of academic verbs, which remains equally suitable over half a year.

9 Results on Extramural English behaviour

In chapter 9.1 findings related to the following research questions will be presented, which are all results that can be related to the 2nd data collection point (T2)⁷⁴, and are therefore not concerned with longitudinal insights.

 *What kinds of extramural English contacts do English major students in Austria report to have?*

- What kinds of recording techniques are used by English major students in Austria in EE settings?
- Which word knowledge components are considered by English major students in Austria in EE settings?
- Is there a connection between biographical background variables and participants' exposure to EE, their recording techniques used and the word knowledge aspects considered in EE settings?

 *How could this EE contact be categorized in a meaningful way based on theoretical and statistical considerations?*

Chapter 9.7 addresses the following research questions, which are concerned with longitudinal results and potential changes that occurred between the 2nd (T2) and the 3rd data collection (T3).

 *Does this exposure to EE change over the course of one academic year?*

- Do these recording techniques change over the course of one academic year?
- Does the attention paid to different word knowledge components in EE settings change over the course of one academic year?
- Which biographical background variables show a connection to EE exposure, recording techniques and word knowledge aspects considered over the course of one academic year and which variables are not stable?

⁷⁴ Since in later chapters the different data collection points are referred to as T1 = 1st data collection, T2= 2nd data collection, T3= 3rd data collection, this labelling is used in this chapter as well for the sake of clarity.

9.1 Involvement in EE activities for T2

Based on data from 153 EE questionnaires, involvement of English major students in 43 EE activities rated via a 5-point Likert-scale was evaluated. Overall engagement in English spare time activities appears to be high, since 97.4% of all participants report to perform at least one of the 43 activities nearly always in English. Only four out of the total 153 participants are not involved in an EE activity nearly always. Overall, a large variety ($IQR=9.5$, $SD=7.273$) regarding the number of EE activities participated in nearly always was found, ranging from 0 to 35 of the named activities. On average participants were performing ten activities in English nearly always and 19 EE activities were participated in at least often.

Regarding the popularity of individual EE activities Figure 9.1 lists the activities which were done nearly always in English by at least 50% of the participants.

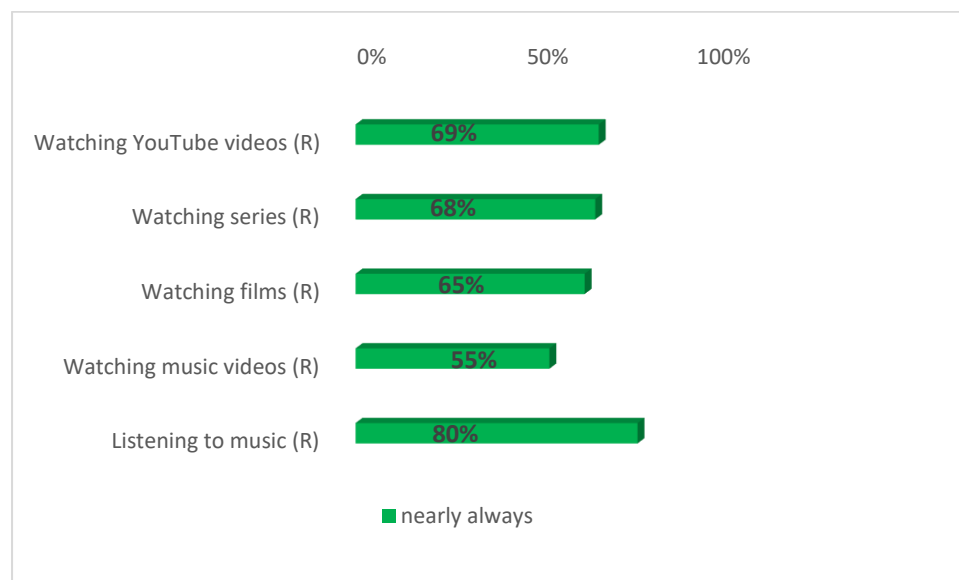


Figure 9.1: EE activities in which at least 50% of the participants are involved in nearly always (N=153)

With 80% of the participants listening to music in English nearly always, the activity is clearly the one most engaged in. The second place goes to watching YouTube videos with no relation to music content with 69% of the participants watching them nearly always in English. This is closely followed by the consumption of English series and English films, with 68% and 65%. Lastly, 55% of the students state that they watch music videos in English nearly always. As shown in brackets next to the activities all five activities performed by over 50% of the participants nearly always are receptive activities. Hence, no action is required on the side of the students. Moreover, four of the named activities are audio-visual activities dependent on screens, which provide participants with both visual and aural input. However, the distribution of visual and aural input can differ dependent on whether participants decide to use subtitles accompanying the spoken language.

Regarding the use of subtitles when consuming audio-visual material, the analysis of open questions of the EE_questionnaire revealed a two-fold division of participants. While 51.95%

of the students appears to scarcely use subtitles (examples 44-46), 32.45% tend to use them rather heavily (examples 47-49).

(44) “I watch series/movies with subtitles about five percent of the time. [...]” (CAB01R)

(45) “Never. Maybe if the movie sound quality is bad.” (AND12VO)

(46) “Overall use of subtitles : Ten percent; subtitles in English: 50%, German: 50%” (CAM02CK)

(47) “I use 99% subtitles in English [...]. I use German subtitles only sporadically” (VID1VO)

(48) “90% of the time. In English. [...]” (MEA11R)

(49) “English subtitles. Almost all the time.” (SIG04SP)

Only 15.58% use subtitles between 40% and 50% of the time. Hence, students appear to have clear preferences concerning the use of subtitles, with the majority of subtitles being in English. Nevertheless, all participants name particular circumstances under which they tend to use subtitles, which can be grouped into four categories as depicted in Figure 9.2.

Movie related parameters	Other languages	Language learning	Surroundings
• pace of speech • clarity of speech • accent used • type of video material	• unknown languages • little proficiency in a language	• knowledge about word form • learning of two languages simultaneously	• social surrounding • local surrounding

Figure 9.2: Reasons for the use of subtitles_EE questionnaire_T2

Firstly, participants report that their use of subtitles is dependent on movie related parameters, such as the pace and clarity of speech, as well as heavy accents of actors. As visible in examples (50) and (51) participants rely on the help of subtitles if actors mumble or speak too fast.

(50) “I use subtitles when actors (person speaking) is/are speaking unclearly. [...]” (AMV02R)

(51) “I use subtitles if the actor is speaking very fast (5%). [...]” (NIC02R)

Moreover, students switch on subtitles if actors have a heavy accent as visible in examples (52)-(55). However, participants appear to have different preferences or perceptions of accent understandability, since they report to use subtitles for different regional dialects.

(52) “I almost never use subtitles. If I do I turn them on in English and because it is a certain dialect in which people are talking.” (CLI02CK)

(53) “I don't use subtitles unless someone uses a strong accent (ex. Scottish)” (ELI05SP)

(54) "If I watch series with British origin I use subtitles. [...]" (THR03VO)

(55) "I only use subtitles when I watch movies where the actors have really strong Australian or Scottish accents which are difficult to understand (10%)" (SOD07CK)

Finally, students reveal to only use subtitles when watching a particular type of video material. For instance, JUM06B mentions that he/she solely uses subtitles when consuming documentaries, while other students report to use them preferably when watching anime. This is directly related to YVI10SP's exemplary statement that "[he/she] only use[s] subtitles if the original content isn't available in English or German". Similarly, many students use subtitles when consuming video material in a language they do not know. However, the use of subtitles is not only restricted to videos in unknown languages but also functions as compensation for insufficient language proficiency in known languages as example (56) highlights.

(56) "I use English subtitles if I watch a film in Spanish. 100% if I watch it Spanish, about 20% of all movies I watch." (LIM04SP)

Nevertheless, it has to be considered that the use of subtitles when watching a movie in an unknown or unfamiliar language might not exclusively be linked to insufficient language knowledge but can be a conscious attempt to learn two languages simultaneously. For example MEA11R states that he/she "use[s] English subtitles to watch anime so [he/she] can learn Japanese and English simultaneously". Opposed to this view VIA07B rejects this approach since he/she "find[s] it awkward to process two languages simultaneously". Therefore, the use of subtitles to improve two languages at the same time appears to depend on personal preferences.

Furthermore, participants describe their use of subtitles to enhance their knowledge about word form as exemplified in (57) and (58).

(57) "I use subtitles often (if so always English). In case I encounter a new word, I already know how to spell it." (NAS03R)

(58) "I use subtitles about 90% of the time because I like to know how certain words are written/spelled." (SOK06B)

Other potential reasons for the use of subtitles named are grounded in the social or local surroundings of participants. As many students indicate that their watching habits are dependent on whether they are alone or in company, family members and partners appear to have a crucial impact on the use of subtitles (see examples 16 & 17). Additionally, disturbance of local surroundings through background noise or eating increase participants use of subtitles (see examples 59-62).

(59) "If I use them-only English- but only if I watch with someone else who doesn't understand everything. Alone never." (KAJ12R)

(60) “I need to use subtitles in German when watching films with my boyfriend, otherwise I do not use subtitles for English films/series at all.” (KAG10VO)

(61) “I watch series/movies with subtitles about five percent of the time. I only watch something with subtitles if: the sound quality is bad, I'm eating something and can't understand the characters [...]”. (CAB01R)

(62) “Sometimes I use English subtitles if my roommates are too loud to hear clearly.” (TUS12R)

Overall, nearly all participants used subtitles at least sporadically. However, it is an interesting observation that participants can indeed be divided into two groups, one using subtitles most of the time and the other who only uses subtitles if required by one of the above mentioned reasons.

When having a closer look at EE activities performed at least often by 50% of the participants, depicted in Figure 9.3 the first three places stay the same.

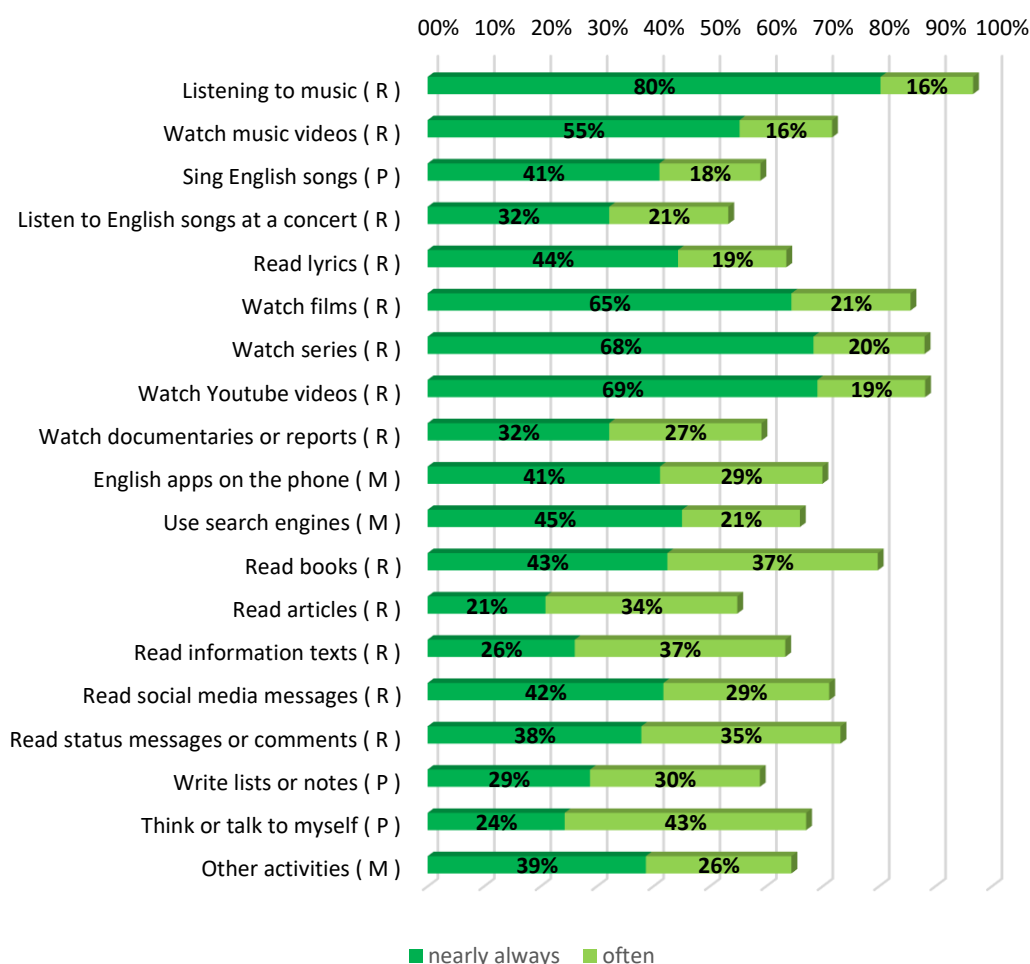


Figure 9.3: EE activities in which at least 50% of the participants are involved in nearly always or often (N=153)

96% of all participants listen to English music at least often and 88% consume English YouTube videos or English series. With 86% watching English films occupies the third place. However, the fourth, fifth and sixth places are taken by reading activities, namely reading English books (80%), reading status messages (73%), and reading social media messages (71%). In comparison genres related to Creative writing, such as fanfiction (*Mdn*=1.0), comics or mangas (*Mdn*=1.0), blogs or forum entries (*Mdn*=1.0) or diaries (*Mdn*=1.0) are rather unpopular overall as Table 9.1 shows. Since all variables are not normally distributed, activities were ordered by the median.⁷⁵ Within the categories activities are ordered by the semi-interquartile range (*IQR*) given in brackets next to each activity, with higher values pointing towards larger within group differences.

Median 5 (N=5)	Listening to music (<i>IQR</i>=0) Watch films (<i>IQR</i>=1) Watch series (<i>IQR</i>=1) Watch YouTube videos (<i>IQR</i>=1) Watch music videos (<i>IQR</i>=2)
Median 4 (N=14)	Think or talk to myself (<i>IQR</i>=1) Read books (<i>IQR</i>=1) Read articles (<i>IQR</i>=1) Read lyrics (<i>IQR</i>=2) Watch documentaries (<i>IQR</i>=2) Use apps (<i>IQR</i>=2) Use search engines (<i>IQR</i>=2) Read information texts (<i>IQR</i>=2) Read social media messages (<i>IQR</i>=2) Read status messages (<i>IQR</i>=2) Write lists/notes (<i>IQR</i>=2) Other activities (<i>IQR</i>=2) Sing songs (<i>IQR</i>=3) Listen to songs at a concert (<i>IQR</i>=3)
Median 3 (N=9)	Speak in face to face interactions (<i>IQR</i>=1) Read short stories (<i>IQR</i>=2) Read blog or forum entries (<i>IQR</i>=2) Read e-mails (<i>IQR</i>=2) Read SMS or WhatsApp messages (<i>IQR</i>=2)

⁷⁵ For information on the descriptive properties (e.g. normal distribution, mean, *SD*) of the individual variables see appendix 9.1.

	Write e-mails (<i>IQR</i> =2) Write social media messages (<i>IQR</i> =2) Write SMS or WhatsApp messages (<i>IQR</i> =2) Write status messages (<i>IQR</i> =3)
Median 2 (N=5)	Write stories (<i>IQR</i> =2) Speak on the phone (<i>IQR</i> =2) Play games on the computer or consoles (<i>IQR</i> =3) Read fanfiction (<i>IQR</i> =3) Speak on Skype or comparable services (<i>IQR</i> =3)
Median 1 (N=8)	Write fanfiction (<i>IQR</i> =0) Compose comics or mangas (<i>IQR</i> =0) Write blogs or forum entries (<i>IQR</i> =1) Write lyrics (<i>IQR</i> =2) Play games online with others (<i>IQR</i> =2) Listen to audiobooks (<i>IQR</i> =2) Read comics (<i>IQR</i> =2) Write diary (<i>IQR</i> =2)

Table 9.1: EE activities sorted by central tendency and interquartile range

While a tendency towards new literacy practices, which incorporate digital genres related to social media, is visible, some classical genres, such as books or articles, remain to be read by a large proportion of participants relatively frequently. Hence, participants, being English major students, seem to be vivid readers targeting a variety of sources overall.

Moreover, Table 9.1 suggests that participants appear to be rather certain about their EE activities not shying away from choosing the extreme ends of the Likert scale. Looking at the interquartile range supports this assumption, since several activities at the extreme ends, with a median of 5 or 1 display the smallest variance. For instance, involvement in very popular activities, such as listening to music (*IQR*=0), watching films (*IQR*=1) or YouTube videos (*IQR*=1), appears to be relatively consistent across the whole group. The same accounts for rather infrequent EE activities, such as writing fanfiction (*IQR*=0), composing comics (*IQR*=0), or writing blog or forum entries (*IQR*=1). Hence, within group variance tends to be stronger for middle categories with a median of 4, 3 or 2, indicating that the group is rather heterogeneous regarding the degree of involvement in these EE activities.

While 12 of the 18 EE activities done often by over 50% of the participants are receptive activities, three productive and three mixed actions were identified. Of the productive activities singing of English songs is most popular with 71% performing English songs at least often. This is followed by thinking or talking to oneself in English, with 67% doing so often and finally writing notes or lists in English, which is done at least by 65% of the participants

often. Concerning activities which require both productive and receptive skills, 70% of the participants use English Apps at least often. Moreover, 66% use search engines, such as Google, in English. Finally, 65% state to participate in other EE activities at least often.

Figure 9.4 visualizes all types of activities named by the participants, with words written larger when named more often by participants.



Figure 9.4: Word cloud for other EE activities named in EE questionnaire_T2

Overall, six participants reported to dream in English at least often, four students stated that they used English during work or when giving tutoring lessons, two people often created art in English and two people often wrote poems in English at least often. All other activities were only named once.

To summarise the main descriptive findings, listening to music and audio visual media, such as watching films or series in English, were the most popular. This is followed by reading books and reading various other digital genres. While receptive activities were clearly preferred by participants, they also engaged in productive activities, such as singing songs, thinking or talking to oneself and writing lists or notes in English.

9.2 Grouping of variables

While most research studies in the field have reported findings on individual variables, similar to the results in section 9.1, this research study has grouped all individual activities listed in Table 9.1 into computational variables according to the theoretical sub-constructs of the EE-questionnaire. Table 9.2 depicts the theoretical sub-constructs of the construct *Extramural behaviour* as used in the EE_questionnaire.⁷⁶ To validate these sub-constructs

⁷⁶ For more information on the EE_questionnaire used for this study see chapter 6.2.2.

for the collected data, principal factor analyses with oblique rotation (direct oblimin)⁷⁷ were run on each of the theoretical sub-constructs listed.

	Sub-constructs
Extramural behavior	English music (EM)
	English video material (EVM)
	English online (EO)
	Reading in English (RIE)
	Writing in English (WIE)
	Speaking in English (SE)

Table 9.2: EE_Questionnaire subconstructs of Extramural behaviour

While the principal factor analyses for the sub-constructs *English music* (EM), *English video material* (EVM) and *Speaking in English* (SE) resulted in only one factor, the constructs *English online* (EO), *Reading in English* (RIE) and *Writing in English* (WIE) had to be refined on the basis of the existent data.⁷⁸ As visible in Figure 9.5 the principal factor analysis for the construct *English online* (EO) resulted in two different factors, indicating a clear differentiation between the variables EO_1 and EO_2 and a second group of variables consisting of EO_3, EO_4, EO_5 and EO_6.

⁷⁷ Rotated factor solutions allow for more differentiated factor-loadings than non-rotated factor analysis, indicating more clearly how strongly variables load on the extracted factors (Loewen & Gonulal, 2015, p.197). In contrast to orthogonal rotation methods, oblique rotation procedures allow factors to be correlated. Since factors are seldom completely independent from each other in social sciences but tend to be related, an oblique rotation method was chosen with direct oblimin, even though varimax might be considered as the default option in exploratory factor analysis (Loewen & Gonulal, 2015, p.197).

⁷⁸ All preconditions for the interpretation of the factor analyses were given. For more detail see appendix 9.2.

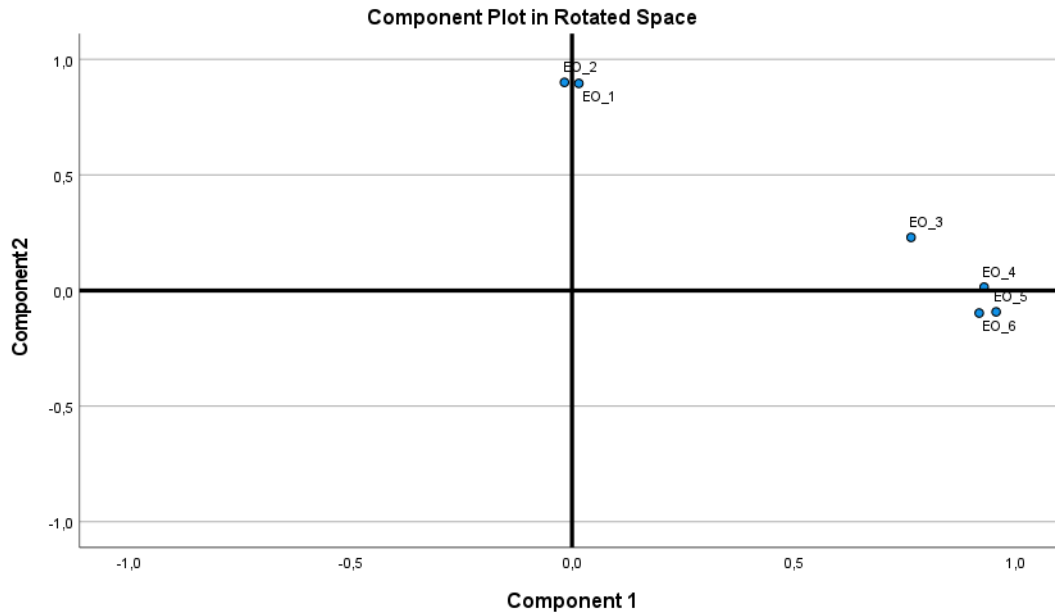


Figure 9.5: 2-factor solution for English online (EO)_EE questionnaire_T2

On the basis of this result the theoretical construct *English online* was split into two constructs, namely *English online* containing EO_1 and EO_2 and *English gaming* EO_3, EO_4, EO_5 and EO_6.

Regarding the construct *Reading in English* (RIE) four factors were originally extracted, with several variables loading strongly on three of them but only *listening to audiobooks* (highlighted in pink) loading strongest on the fourth factor. Hence, it was decided to exclude the variable *listening to audiobooks* from the pool of items for computational variables to reduce the number of factors to 3. After closer investigation it was found that all variables related to *Fiction-oriented reading* (highlighted in green) excluding *reading English comics or mangas*, loaded strongest on the third factor. Variables related to *Information oriented reading* (highlighted in yellow) loaded strongest on the second factor and variables related to *Communication oriented reading* (highlighted in red) on the first factor as visible in Table 9.3. Therefore, the variable *reading English comics or mangas* was excluded and three computational variables were formed on the basis of the existent data, named *Fiction-oriented reading*, *Information oriented reading* and *Communication oriented reading*.

Pattern Matrix^a

	Component			
	1	2	3	4
I read English books.	,175	-,136	,703	,059
I listen to English audiobooks.	-,046	-,044	,136	,936
I read English articles in newspapers, magazines, or news portals.	,193	,463	,081	,106
I read English information texts (instructions, recipes etc.)	,195	,639	-,115	,083
I read English short stories.	-,106	,314	,737	,050
I read English fanfiction.	,030	,048	,739	,020
I read English comics or mangas.	-,212	,701	,092	,042
I read English blogs/ forum entries (about games etc.).	,084	,717	,114	-,292
I read e-mails in English.	,600	,258	-,201	,285
I read social media messages in English.	,892	-,126	,119	-,107
I read status messages or comments in English.	,794	,180	,066	-,089
I read SMS or WhatsApp messages in English.	,756	-,039	,081	,048

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 14 iterations.

Table 9.3: Loading plot 4-factor solution Reading in English (RIE)_EE questionnaire_T2

Concerning the theoretical construct *English writing* three factors were extracted as visible in Table 9.4 below. However, all items but *I write my diary in English* (highlighted in pink) loaded strongest on two of these factors, suggesting a division into two computational sub-constructs, namely *Creative writing* (highlighted in yellow) and *Written communication* (highlighted in red). The item *I write my diary in English* was excluded from the questionnaire as a consequence.

Pattern Matrix^a

	Component		
	1	2	3
I write blogs or forum entries in English.	,042	,669	-,028
I write e-mails in English.	,754	,094	-,301
I write social media messages in English.	,854	-,008	,060
I write status messages or comments in English.	,824	,004	,139
I write SMS or WhatsApp messages in English.	,850	-,033	,022
I write lists or notes in English.	,604	,048	,393
I write my diary in English.	,103	,084	,862
I write English stories.	,156	,564	,284
I write English Fanfiction.	-,143	,833	,197
I compose comics/mangas in English.	,069	,794	-,261

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Table 9.4: Loading plot 3-factor solution Writing in English (WIE)_EE questionnaire_T2

Finally, another factor analysis was run for each of the extracted factors to make sure that each computational variable consists of only one factor and therefore reflects a unified concept. Table 9.5 lists the individual activities subsumed under each computational variable, next to reporting whether the preconditions for interpretation of the factor analysis were given. First, the Kaiser-Meyer-Olkin (KMO) measure is reported, indicating whether sampling adequacy is given (Loewen & Gonulal, 2015, p. 187). “KMO values range from 0 to 1, with higher values representing better sampling adequacy. [...] Values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are perfect.” (Field, 2009, p. 679). Hence, sample size was appropriate for all computational variables. However, even more data would be beneficial for *English video material*, *English online*, *Fiction-oriented reading*, *Information-oriented reading*, and *English speaking*. In a second step it was investigated whether variables are satisfactorily correlated to each other by using Bartlett’s Test of Sphericity. Here “a significant result with $p < .05$ indicates that the variables are correlated and thus suitable for EFA [Exploratory Factor Analysis]”, which is the case for all of the computational variables (Loewen & Gonulal, 2015, p. 188).

Computational variable	Individual activities	Pre-conditions	Interpretation
English music	1. I listen to English music 2. I watch English music videos on the internet 3. I sing English songs (karaoke etc.) 4. I listen to English songs at a concert 5. I read English lyrics 6. I write English lyrics	<i>KMO</i> : 0.812 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.245 <i>Communalities</i> : 0.275-0.609	1 factor which explains 47.052% of the total variance $\alpha = 0.76$
English video material	1. I watch films in English 2. I watch series in English 3. I watch YouTube videos (no music videos) in English 4. I watch documentaries or reports	<i>KMO</i> : 0.684 <i>Bartlett's test of sphericity</i> : < 0.001 <i>R matrix determinant</i> : 0.366 <i>Communalities</i> : 0.408-0.711	1 factor which explains 56.402% of the total variance $\alpha = 0.73$
English online	1. I use English apps on the phone 2. I use search engines such as Google in English	<i>KMO</i> : 0.500 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.602 <i>Communalities</i> : 0.815	1 factor which explains 81.549% of the total variance $\alpha = 0.77$
English gaming	1. Playing games on the computer or on gaming consoles 2. Playing games online with others 3. If I play with others on the internet, I use English in my In-Game chats 4. If I play with others on the Internet, I speak English through	<i>KMO</i> : 0.779 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.030 <i>Communalities</i> : 0.659-880	1 factor which explains 80.406% of the total variance $\alpha = 0.92$

	TeamSpeak, Mumble, or comparable services		
Fiction-oriented reading	1. I read English books 2. I read English short stories 3. I read English fanfiction	<i>KMO</i> : 0.602 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.625 <i>Communalities</i> : 0.505-0.720	1 factor which explains 59.418% of the total variance $\alpha = 0.65$
Information oriented reading	1. I read English articles in newspapers, magazines, or news portals 2. I read English information texts (instructions, recipes etc.) 3. I read English blogs/ forum entries	<i>KMO</i> : 0.636 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.748 <i>Communalities</i> : 0.522-0.568	1 factor which explains 55.155% of the total variance $\alpha = 0.57$
Communication oriented reading	1. I read e-mails in English 2. I read social media messages in English 3. I read status messages or comments in English 4. I read SMS or WhatsApp messages in English	<i>KMO</i> : 0.749 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.236 <i>Communalities</i> : 0.515-0.725	1 factor which explains 63.899% of the total variance $\alpha = 0.81$
Written communication	1. I write e-mails in English 2. I write social media messages in English 3. I write status messages or comments in English 4. I write SMS or WhatsApp messages in English 5. I write lists or notes in	<i>KMO</i> : 0.805 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.074 <i>Communalities</i> : 0.463-0.754	1 factor which explains 64.571% of the total variance. $\alpha = 0.86$

	English		
Creative writing	1. I write English stories 2. I write English Fanfiction 3. I compose comics/mangas in English 4. I write blogs or forum entries in English	<i>KMO</i> : 0.709 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.444 <i>Communalities</i> : 0.463-0.656	1 factor which explains 54,880% of the total variance $\alpha = 0.69$
English speaking	1. I speak English on Skype or other internet services 2. I speak English on the phone 3. I speak English in face to face interactions 4. I think or talk to myself in English	<i>KMO</i> : 0.642 <i>Bartlett's test of sphericity</i> : $p < 0.001$ <i>R matrix determinant</i> : 0.319 <i>Communalities</i> : 0.598-0.834	1 factor which explains 72.205% of the total variance $\alpha = 0.75$

Table 9.5: Information about computational variables EE_questionnaire

Additional to potential non-existent or low correlations between variables, multicollinearity being “the presence of variables that are too highly correlated” needs to be investigated (Loewen & Gonulal, 2015, p. 188). This was approached by examining the determinant of the R-matrix, which should be greater than 0.0001 (Field, 2009). As visible in Table 9.5 multicollinearity is no problem for none of the computational variables. Finally, communalities, “[representing] the amount of common variance in a variable that is accounted for by [...] the extracted [factor]” is considered. Hence, higher communalities for individual variables of a factor are better.

Taking a look at the fourth column of Table 9.5 reveals that for each of the computational variables only one factor was extracted, indicating that each of them represents a self-contained construct. However, the explanatory power indicated by the cumulative percentage of variance varies from construct to construct. While all computational variables but *English music* meet Field’s (2009) threshold of 55-65% explained variance, only four factors adhere to Loewen and Gonulals’ (2015) stricter boundary of over 60%. Hence, explanatory power of *Creative writing*, *Information-oriented reading*, *Fiction-oriented reading*, and *English video material* could be improved even more.

Lastly, Table 9.5 provides information about the reliability of each of the computational variables with the Cronbach α given for each of the factors. While general guidelines (e.g. Kline, 1999) recommend a Cronbach α above the cut-off point of 0.7, Field (2018, p. 823)

demonstrates that Cronbach α is strongly influenced by the number of items the analysed construct comprises, with reliability increasing with the number of items included. Hence, values between 0.5 and 1 can be considered satisfactory for the above tested constructs since they cover only between two and six items.

Based on the analysis outlined above the computational variables *English music*, *English video material*, *English online*, *English gaming*, *Fiction-oriented reading*, *Information-oriented reading*, *Communication-oriented reading*, *Written communication*, *Creative writing*, and *English speaking* will be subsequently used as overarching categories of the associated individual activities. This allows a solid judgement of involvement in different extramural areas and does not restrict the study to estimates of engagement in individual EE activities.

9.3 Differences in EE behaviour regarding biographical background variables

While the information presented in section 9.1 gives an overview of participants' average EE behaviour, it is additionally revealing to see whether biographical background variables examined are related to particular extramural engagement. Overall, the independent variables gender, age of onset, highest educational degree, experience abroad and speaking opportunity in spare time were investigated.⁷⁹ Since normal distribution was not given for all variables non-parametric tests were used for the analysis. For comparisons between two groups the Mann-Whitney U test (Dörnyei, 2007; Field, 2018) was chosen and for more than two groups the Kruskal-Wallis test (Dörnyei, 2007; Field, 2018) was used.

9.3.1 Gender

Regarding participants' gender significant differences could be found for the following computational variables:

<i>English gaming</i>	$U=5.822, z=5.822, p<0.001, r=0.447$
<i>English online</i>	$U=2577.0, z=2.360, p=0.018, r=-0.157$
<i>English music</i>	$U=1438.0, z=-2.495, p=0.013, r=-0.204$
<i>Fiction-oriented reading</i>	$U=1226.5, z=-3.494, p<0.001, r=-0.285$

Table 9.6: Significant differences found for computational variables_gender

In the following these four groups of variables will be examined in more detail to provide a more in depth analysis of the data. Overall, female and male participants differed significantly regarding their engagement in 15 particular EE activities. However, only four activities related to *English gaming* displayed a large effect size between 0.40 and 0.70,

⁷⁹ Moreover, findings on the variables of age and number of languages spoken can be found in appendix 9.1.5. Connections between the biographical background variables and EE engagement are explored both in terms of a relationship with the computational variables and at individual item level.

meaning that gender explained between 40% and 70% of the given variance (Plonsky, 2015, p.38).

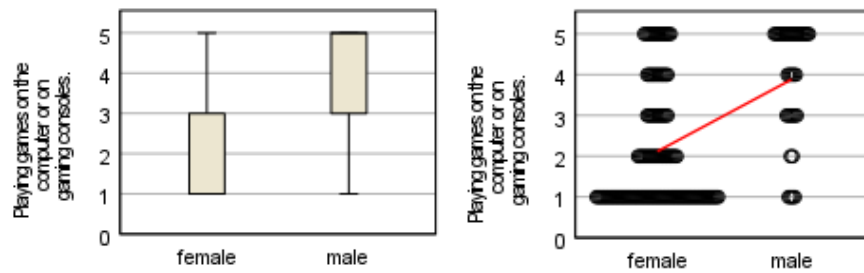


Figure 9.6: Boxplot and Scatterplot with interpolation line_gender & playing computer games

As indicated in the boxplot above males had a considerably higher median ($Mdn=5.0$) than female participants ($Mdn=1.0$) for playing computer games. This descriptive difference proved to be significant with men outperforming women regarding their engagement in gaming on a computer or console ($U=3250.5$, $z=5.528$, $p<0.001$, $r=0.450$). Similar results were found for online gaming ($U=3015.0$, $z=5.059$, $p<0.001$, $r=0.413$). Moreover, an even larger effect was found for the use of In-Game chats ($U=3172.5$, $z=6.078$, $p<0.001$, $r=0.495$) and oral communication services such as Mumble or TeamSpeak, which are used during gaming ($U=3194.0$, $z=6.181$, $p<0.001$, $r=0.503$). Hence, male participants made significantly more use of communication tools when playing games than female participants who played games.

Medium effects were found for engagement in reading English lyrics ($U=1431.5$, $z=-2.735$, $p=0.006$, $r=-0.224$), use of search engines ($U=2633.5$, $z=2.710$, $p=0.007$, $r=0.221$), reading English books ($U=1403.5$, $z=-3.021$, $p=0.003$, $r=-0.246$) and short stories ($U=1423.0$, $z=-2.791$, $p=0.005$, $r=-0.227$) as well as for diary writing ($U=1424.0$, $z=-3.051$, $p=0.002$, $r=-0.250$) (Plonsky, 2015, p. 38).

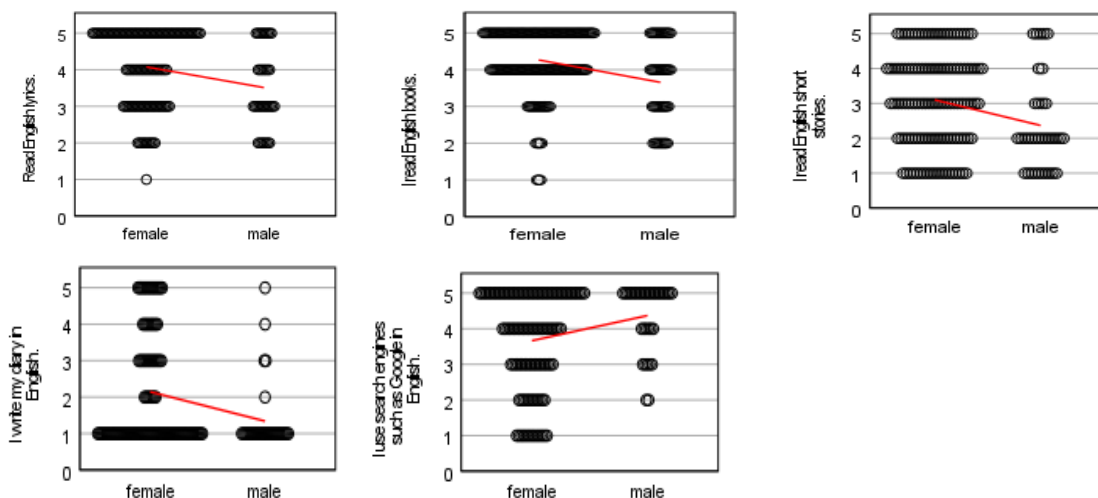


Figure 9.7: Scatterplots with interpolation line for medium effects_gender

As can be seen in the scatterplots above, women engaged more in diary writing as well as in reading English lyrics, books, and short stories. Moreover, women appear to be more involved in writing status messages on social media platforms ($U=1588.5$, $z= -2.058$, $p=0.040$, $r=-0.157$), while men used search engines in English significantly more often. This implies different foci in online usage of English language content. However, the effect was quite small as can be seen in Table 9.6.

Similarly small effects could be found for singing English songs ($U=1535.5$, $z=-2.213$, $p=0.027$, $r=-0.181$) and consuming English music at a concert ($U=1493.5$, $z=-2.433$, $p=0.015$, $r=-0.199$), which corresponds to the significant difference found between male and female participants for consumption of *English music* overall ($U=1438.0$, $z=-2.495$, $p=0.013$, $r=-0.204$). According to these results female students are more involved in English music overall and more particularly in singing English songs and going to concerts.

Additionally, small effects could be found for the overall rather unpopular activities reading ($U=1549.0$, $z=-2.193$, $p=0.028$, $r=-0.180$) and writing of English fanfiction ($U=1732.0$, $z=-2.031$, $p=0.042$, $r=-0.166$), as for the composition of comics or mangas in English ($U=1802.5$, $z=-2.062$, $p=0.039$, $r=-0.168$), with women performing them at least slightly more often than men. This is reflected in the results found for the computational variable *Fiction-oriented reading* ($U=1226.5$, $z=-3.494$, $p<0.001$, $r=-0.285$).

9.3.2 Age of onset

Concerning age of onset participants were divided into the following groups:

0-5 years	24 people
6-9 years	83 people
10-12 years	44 people

Table 9.7: Grouping of participants by age of onset

Between group differences were found for three computational variables (see Table 9.8):

<i>English gaming</i>	$H(2)=20.342$, $p<0.001$
<i>Written communication</i>	$H(2)=11.134$, $p=0.004$
<i>English speaking</i>	$H(2)=10.360$, $p=0.006$

Table 9.8: Significant differences found for computational variables_age of onset

Regarding *English gaming*, people who acquired the English language between the age of 0-5 played significantly more English computer games than participants who started learning English between the age of 10-12 ($H(2)= 3.891$, $p<0.001$, $r= 0.472$). Additionally, a medium effect was found for playing English computer games between an age of onset of 6-9 and 10-12 ($H(2)=2.829$, $p=0.014$, $r= 0.251$), with participants acquiring the language earlier playing more computer games. As visible in Figure 9.8 a Jonckheere-Terpstra test revealed

that there is a significant trend that engagement in playing English computer games decreases the later students have acquired the English language ($\mathcal{F}=2298.0$, $z=-3.977$, $p<0.001$).

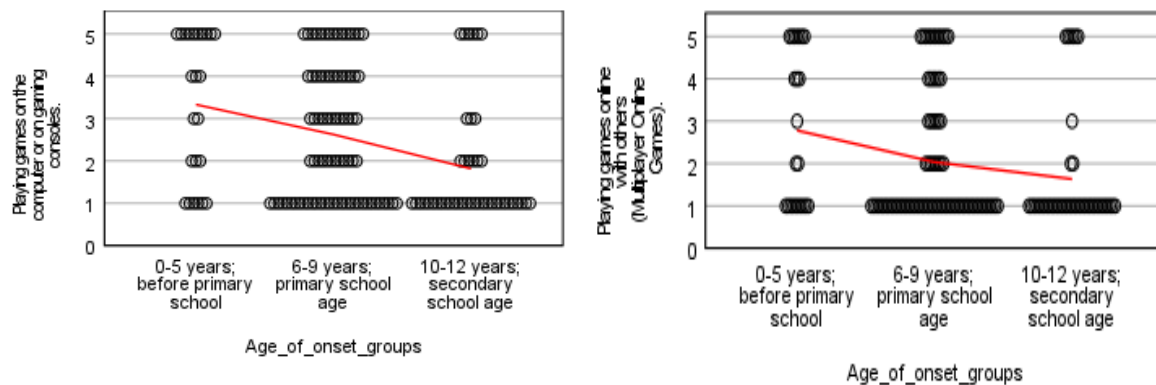


Figure 9.8: Scatterplot with interpolation line_ age of onset & playing computer games offline and online

In connection to playing computer games in English, medium effects were also found for differences regarding playing games online ($H(2)=3.154$, $p=0.005$, $r=0.382$) and the use of In-game chats ($H(2)=2.903$, $p=0.011$, $r=0.352$) and verbal communication channels, such as TeamSpeak ($H(2)=2.792$, $p=0.016$, $r=0.339$). A significant difference was found for all activities between participants aged 0-5 when starting to learn English in comparison to people aged 10-12. These individual differences, observed in all gaming activities, subsume to a large difference for the computational variable *English gaming* overall between the first and the third group ($H(2)=4.464$, $p<0.001$, $r=0.541$) but smaller effects between the other pairings ($H(2)=2.711$, $p=0.020$, $r=0.240$ & $H(2)=2.706$, $p=0.020$, $r=0.262$).

Secondly, a large effect was found for writing: People who acquired the English language between the age of 0-5 wrote significantly more SMS or WhatsApp messages in English than participants who acquired English during secondary school ($H(2)=3.409$, $p=0.002$, $r=0.423$). Again a medium effect is detectable for age of onset for 6-9 and 0-5 ($H(2)=3.409$, $p=0.002$, $r=0.325$) but not between 6-9 and 10-12 ($H(2)=0.468$, $p=1.00$). Hence, the amount of SMS or WhatsApp messages written in English (see Figure 9.9), decreased the later the English language was acquired ($\mathcal{F}=2605.5$, $z=-2.748$, $p=0.006$).

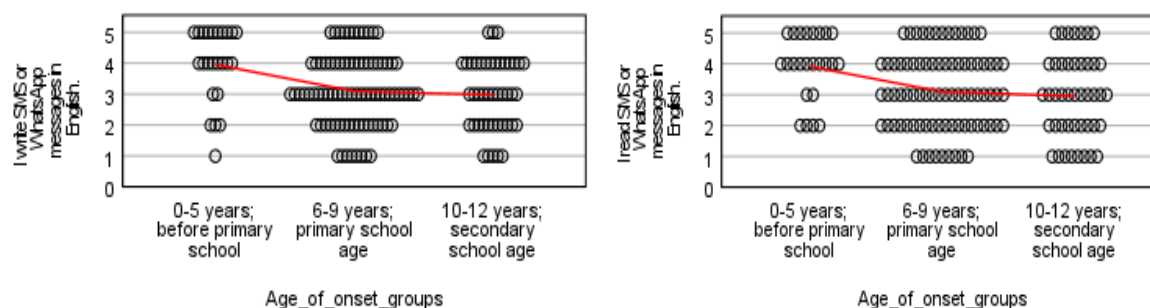


Figure 9.9: Scatterplot with interpolation line_age of onset & writing and reading SMS or WhatsApp messages

Comparable results were found for the composition of e-mails in the English language ($H(2)=2.786, p=0.016, r=0.338$). Overall, the differences regarding involvement in writing SMS or WhatsApp messages and e-mails can be linked to the observed medium effects in the computational variable *Written communication* between 0-5 year olds and 6-9 year olds ($r=0.389$) and 10-12 year olds ($r=0.280$).

An identical trend has been found for speaking English on the phone, with the age of onset group between 0-5 outperforming both later groups, 6-9 ($H(2)=2.810, p=0.015, r=0.272$) and 10-12 ($H(2)=3.904, p<0.001, r=0.473$). As graphically depicted in Figure 9.10, a steady decrease of English speaking time on the phone is visible the older the participants were when learning English ($\mathcal{F}=2311.5, z=-3.659, p<0.001$).

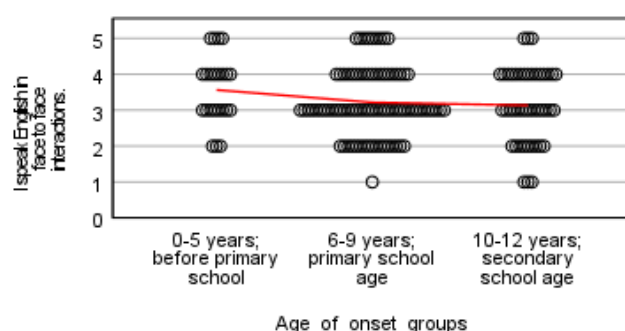


Figure 9.10: Scatterplot with interpolation line_age of onset & speaking English in face to face interactions

Moreover, medium effects were found for other activities related to *English speaking*. Similar to the engagement in speaking English on the phone, participants acquiring English the earliest outperform participants who started learning English in secondary school in the frequency with which they speak English via Skype ($H(2)=2.785, p=0.021, r=0.338$).

Finally, differences were found concerning activities related to *Creative writing* and *Communication oriented reading*. Again participants acquiring the language earlier outperformed groups starting later. Firstly, participants aged 0-5 when learning the English language write more stories in English than participants aged 10-12 ($H(2)=2.789, p=0.016, r=0.339$). Secondly, this group of participants also spends more time on composing English

comics in comparison to the groups aged 6-9 ($H(2)=2.950$, $p=0.011$, $r=0.285$) and 10-12 ($H(2)=2.916$, $p=0.010$, $r=0.354$). Regarding noting down English lyrics the group, which started learning English in secondary school invests less time than both other groups ($H(2)=2.849$, $p=0.013$, $r=0.253$ & $H(2)=2.872$, $p=0.012$, $r=0.348$). For the receptive activity reading of SMS or WhatsApp messages in English a medium effect was detected between the participants who acquired English the earliest and the two other groups, 6-9 ($H(2)=2.958$, $p=0.009$, $r=0.286$) and 10-12 ($H(2)=3.096$, $p=0.006$, $r=0.375$). Regarding reading of e-mails ($H(2)=2.658$, $p=0.024$, $r=0.322$) the first group, which was aged between 0-5 when learning English, outperformed the last group, aged 10-12. However, all differences within these individual variables were not strong enough to yield significant results at the level of computational variables. In other words no significant differences were found for *Creative writing* and *Communication oriented reading* overall.

9.3.3 Highest educational degree

Regarding their highest educational degree 111 participants were graduates from a secondary school and 33 from a vocational school.⁸⁰ Only five people had received a BA degree by the time of observation and only three people a MA degree.⁸¹

Differences have been detected for students who had just graduated from school and students who had already received a BA degree. Overall, nine large effects could be identified, all but one concerning activities closely linked to communication:

Communication oriented reading	$H(4)= 13.361$, $p=0.010$
Written communication	$H(4)= 13.828$, $p=0.008$
English speaking	$H(4)=13.501$, $p=0.009$

Table 9.9: Significant differences found for computational variables_highest educational degree

Regarding reading in English significant differences were found for the frequency of reading e-mails ($H(4)= 10.224$, $p=0.037$), social media messages ($H(4)= 9.949$, $p=0.041$) and SMS or WhatsApp messages ($H(4)= 11.580$, $p=0.021$). While medium effects were found between secondary school graduates and participants with a BA for reading e-mails ($H(4)= -3.070$, $p=0.021$, $r=-0.258$) and SMS or WhatsApp messages in English ($H(4)=-2.907$, $p=0.036$, $r=-0.270$) large effects were found for vocational school graduates and people with a BA for all three activities ($H(4)=-2.939$, $p=0.033$, $r=-0.477$; $H(4)=-2.778$, $p=0.055$, $r=-0.451$; $H(4)= -2.867$, $p=0.041$, $r=-0.465$). These results indicate that participants holding a BA degree spend

⁸⁰ While secondary schools in Austria have the main aim of conveying a thorough general education in eight years, vocational colleges comprise five years of first vocational additional to general education starting from the ninth grade onwards.

⁸¹ While the results are valid according to the current standards of the field, attention should be drawn to the fact that different group sizes can lead to disturbed results in rank based tests, such as the Kruskal-Wallis test, as outlined in Brunner et al. (2018). However, such phenomena are still under research with differing opinions and practices in the field.

significantly more time on reading e-mails, social media messages and SMS or WhatsApp messages in English than graduates from both school types. These differences are subsumed in the category of *Communication-oriented reading*, for which a medium difference was found between secondary school graduates and students with a BA ($H(4)=-3.354$, $p=0.008$, $r=-0.311$) and a large effect for vocational school students in comparison to participants with a BA ($H(4)=-3.533$, $p=0.004$, $r=-0.573$).

Similar results could be detected for writing activities in English related to communication practices. Differences between people with a school leaving certification and participants holding a BA degree were found for writing e-mails ($H(4)= 17.365$, $p=0.002$), social media messages ($H(4)= 10.994$, $p=0.027$), status messages ($H(4)= 10.067$, $p=0.039$) and SMS or WhatsApp messages in English ($H(4)= 12.595$, $p=0.013$). While for writing e-mails ($H(4)=-3.548$, $p=0.004$, $r=-0.329$), social media messages ($H(4)=-3.019$, $p=0.025$, $r=-0.280$) and status messages ($H(4)=-2.805$, $p=0.050$, $r=0.260$) medium effects were found between secondary school graduates and participants with a BA degree, large effects were found for vocational school graduates and participants with a BA degree for all four written activities.

Hence, similar to the corresponding receptive actions participants with a BA degree outperform students holding a school leaving certification in the production of e-mails, social media messages, status messages and SMS or WhatsApp messages in English. As can be seen in Figure 9.11 the largest difference between the groups was found for writing e-mails in English ($H(4)=-3.875$, $p=0.001$, $r=-0.629$), followed by social media messages ($H(4)=-3.088$, $p=0.020$, $r=-0.501$), SMS or WhatsApp messages ($H(4)=-3.013$, $p=0.026$, $r=-0.489$) and status messages ($H(4)=-3.006$, $p=0.027$, $r=0.488$). Again the same effects were found in the overarching variable *Written communication* ($H(4)=-3.391$, $p=0.007$, $r=-0.315$; $H(4)=-3.696$, $p=0.002$, $r=-0.510$).

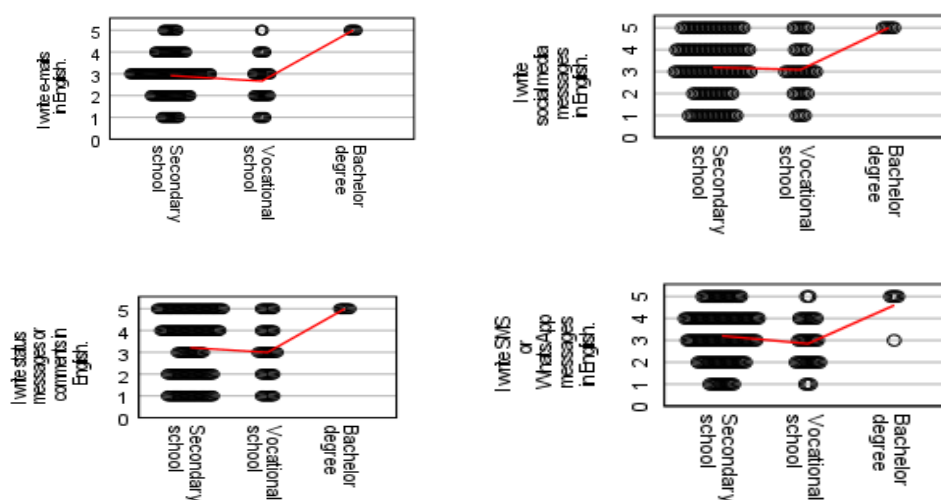


Figure 9.11: Scatterplots with interpolation lines_highest educational degree & written communication activities

Finally, differences between secondary school graduates or vocational school graduates and participants with a BA degree were detected for *English speaking* activities ($H(4)=12.236$, $p=0.016$). Participants with a BA appear to be talking more English on the phone ($H(4)=13.501$, $p=0.009$) and in face to face interactions ($H(4)=15.414$, $p=0.004$) in comparison to participants who have just left school. Again a medium effect was detected for the difference between BA and secondary school graduates ($r=-0.319$; $r=-0.312$) and a large effect for vocational school graduates ($r=-0.570$; $r=-0.522$) for both activities.

9.3.4 Experience abroad

Similar to the biographical variable of education, experience abroad made a difference regarding EE practices linked to *Written communication in English*. As visible in Figure 9.12, participants who had stayed abroad appear to read more SMS or Whatsapp messages in English ($U=2577.0$, $z=2.074$, $p=0.038$, $r=0.169$) and write more e-mails in English ($U=2632.0$, $z=2.261$, $p=0.024$, $r=0.189$) than participants who have no experience abroad.

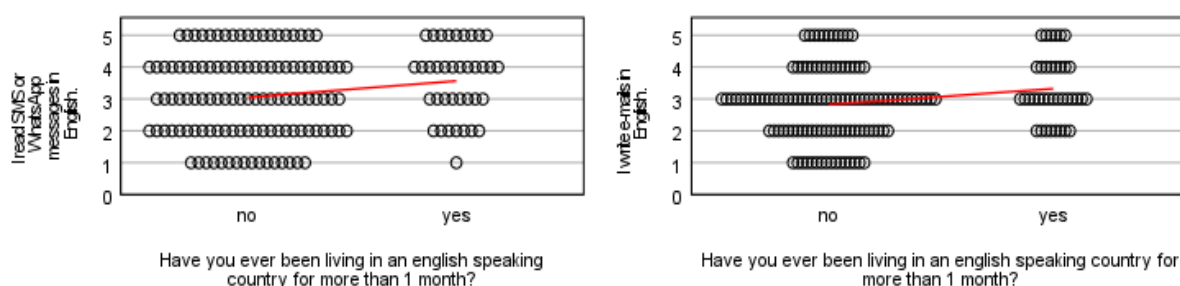


Figure 9.12: Scatterplots with interpolation lines_experience abroad & reading SMS or WhatsApp messages & writing e-mails

Moreover, participants with experience abroad speak more English in their spare time overall ($U=2718$, $z=2.635$, $p=0.008$, $r=0.228$), with significant differences found for speaking English via Skype or comparable services ($U=2763.0$, $z=2.917$, $p=0.004$, $r=0.238$) and on the phone ($U=2737.5$, $z=2.809$, $p=0.005$, $r=0.229$).

9.3.5 Speaking opportunity in spare time

Regarding speaking opportunities in their spare time 101 participants stated that they regularly speak English outside of a professional or educational setting. However, at least 52 participants report that they have no opportunity to speak English in their free time, which appears to be a high number for English major students. What is, however, even more crucial is that the fact whether participants regularly speak English in their spare time had a huge influence on their overall EE behaviour, with participants speaking English regularly outperforming participants with no regular spoken contact. Overall, the two groups differed significantly in five computational variables and 21 individual extramural activities:

English online	$U=3532.5, z=3.559, p<0.001$
Communication oriented reading	$U=3575.5, z=3.904, p<0.001$
Creative writing	$U=3322.0, z=2.982, p=0.003$
Written communication	$U=3962.0, z=5.149, p<0.001$
English speaking	$U=4008.5, z=5.471, p<0.001$

Table 9.10: Significant differences found for computational variables_speaking opportunity

Firstly, students with regular speaking opportunities are more engaged in *English online* activities ($U=3532.5, z=3.559, p<0.001$), such as the use of English Apps ($U=3367.5, z=3.009, p=0.003, r=0.244$) and the use of search engines, such as Google ($U=3413, z=3.206, p=0.001, r=0.260$). Another online activity, however linked to *English gaming*, which is more practiced by participants who have spare time speaking opportunities is playing games online with other people ($U=3124.0, z=2.498, p=0.013, r=0.203$).

Furthermore, many differences between the two groups can be found for reading practices of different genres. In addition, to reading more English articles ($U=3296.0, z=2.713, p=0.007, r=0.220$) and English short stories ($U=3195.5, z=2.240, p=0.025, r=0.182$), students who speak English regularly in their spare time also spend more time on reading e-mails ($U=3265.0, z=2.539, p=0.011, r=0.206$), social media messages ($U=3490.5, z=3.516, p<0.001, r=0.285$), status messages ($U=3419, z=3.224, p=0.001, r=0.245$) and SMS or WhatsApp messages ($U=3530.5, z=3.817, p<0.001, r=0.220$) in English. Hence, participants with more speaking opportunities were more frequently engaged in *Communication-oriented reading* practices ($U=3575.5, z=3.904, p<0.001$) as can be seen in Figure 9.13.

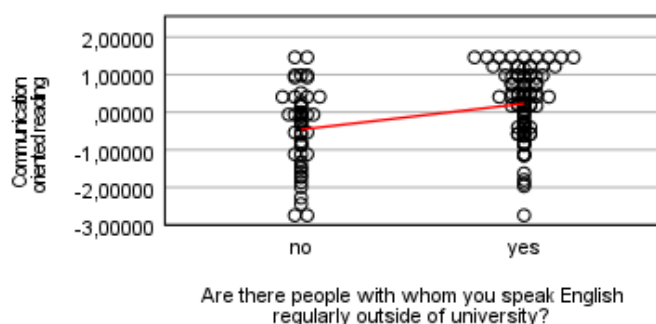


Figure 9.13: Scatterplot with interpolation line_speaking opportunities & communication-oriented reading

Moreover, they additionally outperform participants with no speaking opportunities in their spare time concerning the production of various different genres. Again the group with spare time speaking opportunities is more involved in *Creative writing* ($U=3322.0, z=2.982, p=0.003$), including writing English stories ($U=3337.0, z=3.013, p=0.003, r=0.325$) or comics/mangas ($U=2938.0, z=2.707, p=0.007, r=0.262$) and invests more time in *Written communication* ($U=3962, z=5.149, p<0.001$) as depicted in Figure 9.14. More time is spent on writing e-mails ($U=3623.5, z=4.008, p<0.001, r=0.311$), social media messages ($U=3928.5,$

$z=5.137$, $p<0.001$, $r=0.417$), status messages ($U=3.681.0$, $z=4.162$, $p<0.001$, $r=0.338$), SMS or WhatsApp messages ($U=3718.0$, $z=4.322$, $p<0.001$, $r=0.351$) and lists or notes in English ($U=3371.0$, $z=2.963$, $p=0.003$, $r=0.240$).

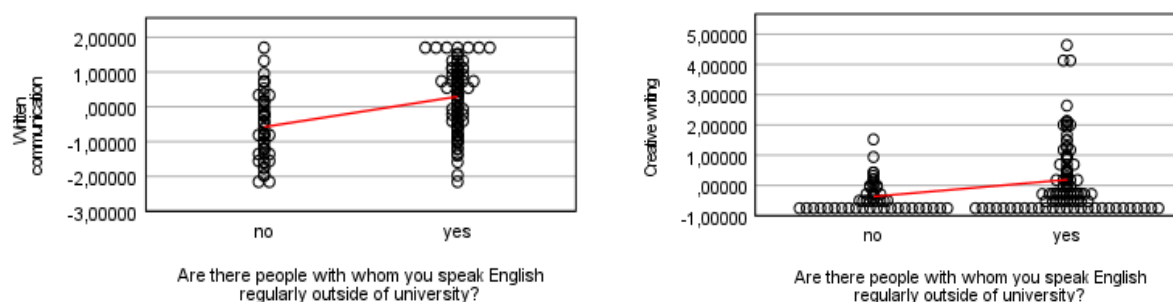


Figure 9.14: Scatterplots with interpolation lines_speaking opportunities & written communication & Creative writing

Finally, a difference regarding EE activities related to *English speaking* ($U=4008.5$, $z=5.471$, $p<0.001$, $r=0.444$) has been found, which supports the correctness of the information provided by the students (see Figure 9.15).

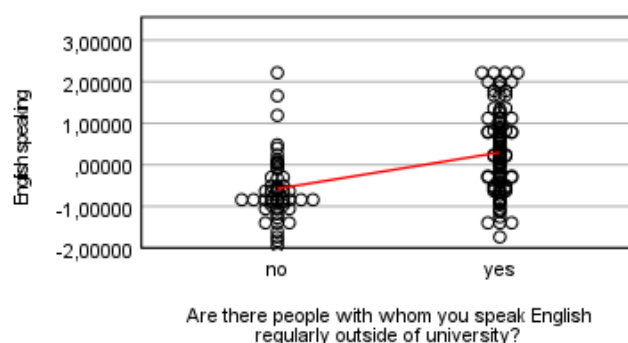


Figure 9.15: Scatterplot with interpolation line_speaking opportunities & English speaking

Significant differences could be found for all four English speaking activities asked for. The largest differences between the groups was found for speaking English via Skype or comparable services ($U=3037.5$, $z=5.358$, $p<0.001$, $r=0.436$), followed by speaking English on the phone ($U=3819.0$, $z=4.888$, $p<0.001$, $r=0.398$) and face to face interactions ($U=3337.0$, $z=3.000$, $p=0.003$, $r=0.244$). This corresponds to the information given by participants regarding their communication partners, since many students report to be in contact with English speaking people abroad, such as former host families, friends, family members living abroad or online acquaintances. However, some participants speak English regularly with partners or flatmates, which will most likely happen in a face to face setting. Only a small difference was found between the groups regarding the frequency in which they thought or talked to themselves in English ($U=3124.5$, $z=2.152$, $p=0.031$, $r=0.175$).

Short summary of the main findings

Overall, differences were found for all five biographical variables investigated, ranging from large to small effects. Concerning gender, men appear to be more involved in *English gaming*, while women practice more *English online*, listen more to *English music* and prefer *Fiction-oriented reading*. Moreover, it was found that the earlier participants acquired the English language the more they were involved in activities related to *English gaming*, *Written communication* and *English speaking*. A similar effect was detected for education, since participants holding a first university degree were more involved in *Communication oriented reading*, *Written communication* and *English speaking* than school graduates. Yet, the effect was consistently stronger between students holding a BA degree and vocational school graduates than for secondary school graduates. Lastly, participants with more speaking opportunities in their spare time use more *English online* and were more involved in *Communication oriented reading*, *Written communication*, *Creative writing* and *English speaking*.

9.4 Attention to new words

When examining all variables related to the type and degree of conscious attention that participants pay to new words in their spare time, it becomes visible that while a focus on the language is no prerequisite for extramural English learning to take place, most participants appear to pay attention to new vocabulary items if they are relevant for the content ($Mdn=4.0$, $IQR=1.0$). Only 0.7% claim to ignore new words even if they are necessary to understand the content. This corresponds to the finding that participants rather seldom recognized a new word they have not paid conscious attention to ($Mdn=2.0$, $IQR=1.0$) or are not interested in new words in their free time at all ($Mdn=1.0$; $IQR=1.0$). In particular students appear to take note of new words if encountered in new media ($Mdn=4.0$, $IQR=1.0$). At least 14.5% of the participants say they always pay attention to new items in new media, followed by 38.2%, who state they do so at least often. Again a very small percentage of 1.3% reports to do so nearly never. Hence, overall participants appear to be paying conscious attention to new words encountered during their EE activities relatively frequently.

9.5 Recording techniques

Although many participants pay attention to new words in their spare time, participants deal differently with these words. While only two percent of the participants state that they tend to do nothing when recognising a new word, most participants reject this claim ($Mdn=1.0$, $IQR=2.0$). Instead participants rather write down new words in a list ($Mdn=3.0$, $IQR=2.0$) or in a rather unorganized manner ($Mdn=3.0$, $IQR=2.0$). More precisely, participants vary considerably more regarding their preferred recording techniques than concerning the attention to new words. While 14.5% report that they always write down

new words in a list and 27.6% state that they at least do so often, 21.7% declare that they never record words in a list. Similarly, 42.7% state that they nearly always or often write down new words in an unorganized form, but 33.5% state to never or seldom record words in that manner. Moreover, a similar number of people reports that they do not write down new words at all but that they simply repeat them orally ($Mdn=3.0$; $IQR=2.0$). Therefore, participants appear to vary considerably regarding their preferred recording techniques. However, more elaborate recording techniques, such as making flash cards or keeping a private vocabulary folder, are not as popular ($Mdn=1.0$) than repeating words orally, writing them down in a list or in an unorganized manner.

Looking at the correlation matrix in Figure 9.16 it becomes visible that participants appear to apply several of the recording techniques mentioned. For instance, students who keep a word list also tend to keep a vocabulary booklet ($r_s=0.492$, $p<0.001$). Interestingly, keeping a word list does also correlate strongly with repeating the words orally instead of writing them down in whatever form ($r_s=0.412$, $p<0.001$) and doing nothing ($r_s=0.518$, $p<0.001$). Hence, the participants might be following several strategies, sometimes writing down new words in a word list or vocabulary booklets but sometimes also repeating words orally or doing nothing.

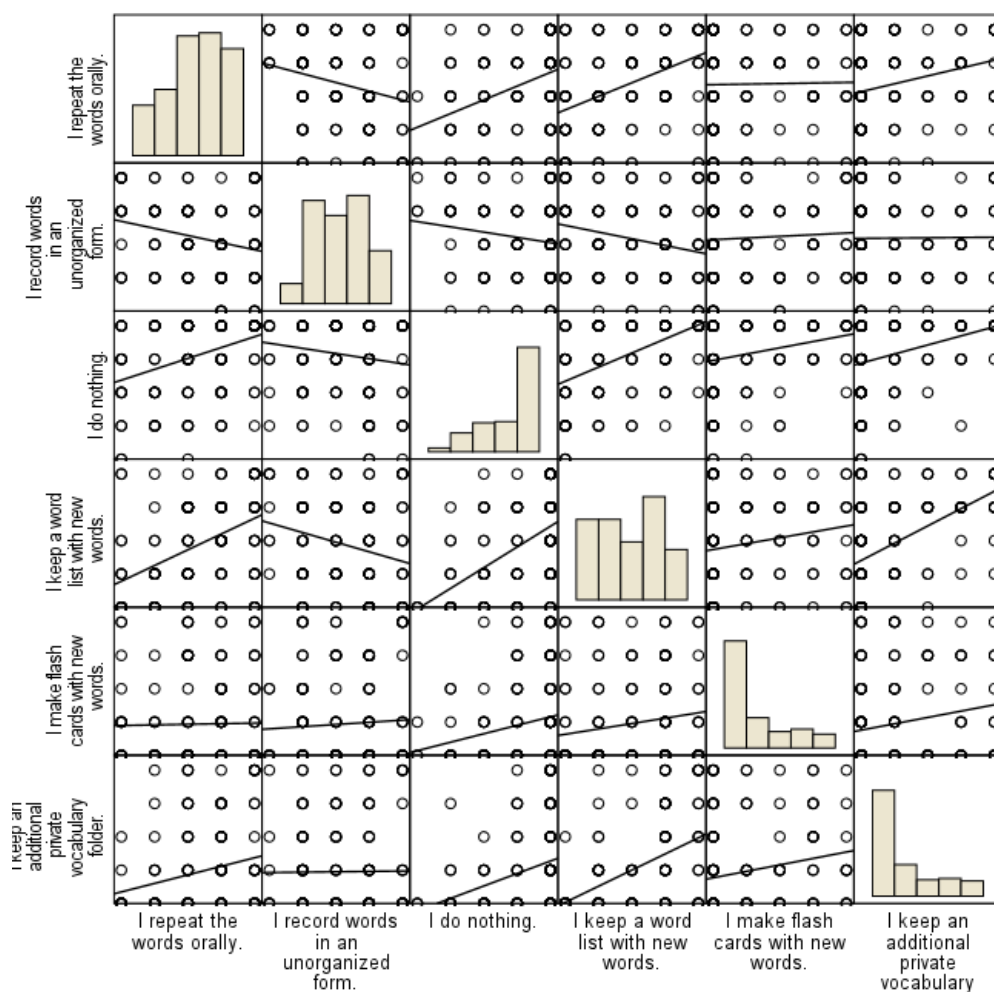


Figure 9.16: Correlation matrix of recording techniques used in an EE setting_T2⁸²

Additionally, differences regarding the recording behaviour of participants were found for the biographical variables of gender and highest educational degree. Firstly, the recording strategy of keeping a private vocabulary folder appears to become more popular with higher educational degrees, with BA students surpassing vocational school graduates ($H(4)=-3.619$, $p=0.003$, $r=-0.587$). Secondly, statistically significant differences were found between male and female participants for repeating the words orally ($U=1306.5$, $z=-3.069$, $p=0.002$, $r=-0.251$) and doing nothing ($U=1458.0$, $z=-2.433$, $p=0.015$, $r=-0.201$), with a larger proportion of female students practicing these recording techniques.

9.6 Word knowledge aspects

While it is clear by now that participants tend to pay conscious attention to new words in their spare time without spending too much effort on recording techniques such as

⁸² The correlation matrix depicts the relationship between two variables. If the line is rising, the relationship is direct (e.g. the more the more) and if the line is falling the relationship is indirect (e.g. the more the less). The steepness of the line indicates how strong the detected connection between two lines is. If the line is flat, such as for the relation between *I keep a private vocabulary folder* and *I record words in an unorganized form*, this means that no relationship can be detected.

vocabulary folders or flash cards, this section is going to reveal which word knowledge aspects participants consider when encountering a new word in their free time. All word knowledge aspects considered are based on Nation's taxonomy (2013) (see chapter 2.4.2.1).

As visible in Table 9.11 participants consider only a particular number of word knowledge aspects regularly in their free time, while many aspects are examined rather rarely. Moreover, students appear to avoid extremes, since the median was five only once and a median of 1 does not occur at all. This stands in contrast to participants' rating behaviour regarding attention to word knowledge or extramural behaviour, since poles were not avoided for these constructs.

Median 5.0	1. I look up definitions. (Form-meaning knowledge)
Median 4.0	1. I consider the spelling of a new word. (Written form) 2. I repeat a new word orally and practice pronunciation. (Spoken form) 3. I try to find words in other languages that help me. (Word parts) 4. I search for translations. (Form-meaning knowledge) 5. I look up in which different contexts the word is used. (Concepts and referents) 6. I search for synonyms. (Associations)
Median 3.0	1. I look for different shades of meaning of a word. (Concepts and referents) 2. I search for antonyms. (Associations) 3. I use the word in a sample sentence. (Grammatical functions) 4. I look up the correct grammatical usage of the word. (Grammatical functions) 5. I look for frequent collocations of the word. (Collocations)
Median 2.0	1. I look up the phonetic transcription of a new word. (Spoken form) 2. I try to divide the word into word parts that I understand. (Word parts) 3. I look up other members of the word family. (Word parts) 4. I search for the word class of the word. (Grammatical functions) 5. I look up the level of formality of a word. (Constraints on use) 6. I look up the currency of the word in modern language use. (Constraints on use) 7. I look up whether the word has meaning differences in different varieties of English such as British English or American English.

	(Constraints on use)
	8. I think about the level of abstraction of the word. (Constraints on use)
	9. I look up whether the word is restricted to a specific domain, such as law or medicine. (Word function)

Table 9.11: Frequency of consideration of word knowledge aspects in free time sorted by median

Considering Nation's taxonomy (2013), one can say that aspects related to form-meaning knowledge of a word, namely definitions ($Mdn=5.0$, $IQR=1.0$) and translations ($Mdn=4.0$, $IQR=1.0$), are researched very often by most participants, since little variance within the group can be detected. Moreover, two more aspects related to knowledge about the meaning of a word appear to be popular. Firstly, participants seem to be interested in the different contexts in which a new word can be used ($Mdn=4.0$, $IQR=2.0$) and secondly, participants report to search for synonyms for new words ($Mdn=4.0$, $IQR=2.0$). However, there is a little more in group variance for these two slightly more advanced activities than for the relatively basic actions of searching for a translation or a definition. Another aspect worth mentioning is that the group of participants tends to try to relate new words to other languages ($Mdn=4.0$, $IQR=3.0$). Nevertheless, one needs to consider that there is high variance, with some people reporting to link new words to other languages very frequently and others answering that they do so very rarely. Finally, participants frequently consider the productive written ($Mdn= 4.0$, $IQR= 1.0$) and spoken form ($Mdn=4.0$, $IQR=2.0$) of new words, which indicates that they attribute importance to the production of new words for written texts or spoken communication.

While one could assume that participants will vary considerably in their consideration of word knowledge aspects assigned to the middle category, this is not the case for the two remaining aspects related to word meaning. Hence, most participants appear to search for different meaning shades ($Mdn=3.0$, $IQR= 1.0$) and antonyms ($Mdn=3.0$, $IQR=1.0$) from time to time. The same applies to information about the correct grammatical use of a word ($Mdn=3.0$, $IQR=2.0$) and collocations ($Mdn=3.0$, $IQR=2.0$); yet with slightly more in group variance than for the two word aspects related to meaning. Nevertheless, most variance has been found for the use of sample sentences ($Mdn=3.0$, $IQR=3$) in this category.

While most participants appear to agree on considering word knowledge aspects related to word meaning in their free time, a more divergent picture is found for aspects related to word function and word form. Although consideration of productive word knowledge aspects related to form is quite popular, participants seldom base their oral production on researching phonetic transcriptions of words ($Mdn=2.0$, $IQR=2.0$). Moreover, while participants are rather decided about not considering other word family members of a new word ($Mdn=2.0$, $IQR=1.0$), there are large differences between group members regarding their preference of examining word parts ($Mdn=2.0$, $IQR=3.0$), although this word aspect is also not considered very often overall.

Regarding word aspects related to word function it becomes clear that participants seldom consider aspects related to constraints of use of a word when they encounter a new word in their spare time. Neither formality ($Mdn=2.0$, $IQR=2.0$), currency ($Mdn=2.0$, $IQR=2.0$) or level of abstraction ($Mdn=2.0$, $IQR=2.0$) of a word are taken into account regularly. Similarly, participants are not interested in meaning or usage differences of words caused by different varieties ($Mdn=2.0$, $IQR=2.0$) or domains ($Mdn=2.0$, $IQR=1.0$) in their spare time.

While similarities and differences between participants can already be detected by examining the interquartile range of the non-normally distributed data, some statistically significant differences have been found for *gender*, *existence of speaking opportunities*, *highest educational degree* and *age*.

Firstly, females appear to outperform male participants regarding the frequency of consideration of synonyms ($U=1336.0$, $z=-3.190$, $p=0.001$, $r=-0.260$) and different shades of meaning ($U=1595.0$, $z=-1.980$, $p=0.048$, $r=-0.162$) of new words. While the effects are rather small, females appear to search for synonyms of words encountered in their spare time more often and they think about possible meaning shades of these words more frequently.

Regarding the existence of speaking opportunities, a difference was found for paying attention to different members of a word family ($U=2095.5$, $z=2.032$, $p=0.042$, $r=0.165$). Students, who speak English regularly in their spare time appear to be more engaged in searching for different word family members of a newly encountered word in their free time, than participants who have no opportunity to speak English regularly. Another difference related to the knowledge of word form has been found for two different age groups, with 24-26 year old participants spending significantly more time on analysing word parts of new items than 20-21 year olds ($H(4)=-3.227$, $p=0.012$, $r=-0.375$). Similarly, the handling of word parts differs between secondary school graduates and students holding a BA degree ($H(4)=-2.802$, $p=0.051$, $r=-0.260$). As the results show BA students appear to be more aware of word parts than secondary school graduates and accordingly examine them more often in their free time. Interestingly, the reverse picture is found for another word knowledge aspect related to word meaning, namely definitions. Closer examination reveals that older participants aged 26 or above consider definitions less frequently than younger participants aged between 22 and 23 ($H(4)=2.893$, $p=0.038$, $r=0.476$). Hence, it might be argued that consideration of different word knowledge aspects changes over time on basis of participants' changing awareness and needs.

9.7 Involvement in EE activities T3

While at the 2nd data collection 153 participants filled in the EE_questionnaire only 110 of these participants provided the questionnaire at the 3rd data collection. This equates a drop-out rate of 28%, which is, considering the temporal distance of the two data collection points, a reasonable number. However, in order to allow for a direct comparison between

participants' involvement in EE activities it has to be investigated whether reducing the original participant number of 153 at T2 to the 110 participants left at T3 will make a significant difference. Closer examination of the data after deleting 43 of the original group of 153 participants has revealed no large differences regarding any of the variables.⁸³ Hence, the data collected at T3 will be restricted to the same 110 participants as found for T2, altering the percentages given for the involvement in EE activities only slightly.

Overall, engagement with EE activities appears to be consistent, since no significant differences were found for the 10 computational variables and only three significant differences for individual variables were found between the two points in time. While listening to English music was the most popular EE activity at T2, participants appear to consume less English music at T3 ($U=38.0$, $z=-2.744$, $p=0.006$, $r=-0.259$). Moreover, participants appear to read English books more seldom. While this activity was quite popular at T2, being in fourth place of activities performed at least often by over 50% of the participants, students appear to spend significantly less time on reading English books at T3 ($U=496.5$, $z=-2.303$, $p=0.021$, $r=-0.218$). This means that activities related to *Communication oriented reading*, such as reading social media messages and status messages or comments in English, have become more popular than reading books. Conversely to music and book consumption watching YouTube videos in English increased significantly ($U=286.0$, $z=-3.210$, $p=0.043$, $r=0.192$). This involves that watching YouTube videos has overtaken listening to music in the ranking of activities which were nearly always performed in English, making it the most popular EE activity at T3.

Overall, audio-visual media remained very popular with comparable findings regarding the use of English subtitles. Similar to T2 popularity of subtitles is again two-fold, with 50.79% of the participants using few subtitles and 36.51% drawing on them rather heavily. Examples (63) and (64) are examples of students who never use English subtitles, while (65) and (66) illustrate a heavy use of subtitles.

(63) "Ten percent; I try also to watch films in French (my second subject) with French subtitles." (MAG10RO)

(64) "never (0%)" (LAB08B)

(65) "I use English subtitles 100% of the time." (LEA01SP)

(66) "I use English subtitles 90% of the time, but I only actively read them occasionally." (ANP01CK)

Only 12.70% of the participants used subtitles between 39% and 59%, indicating a neutral result of subtitling and original media. Interestingly, participants also appear to differ in their type of subtitle consumption. While ANP01CK in (66) suggests a largely passive treatment of subtitles, JAU11SP and FEV07SP indicate in (67) and (68) that they feel compelled to consume subtitles actively, which they perceive as distracting experience.

⁸³ For more detailed information on the impact of the drop-out rate on the data distribution see chapter 6.3.

(67) “I (almost) never watch movies/series with subtitles as they distract me from what is happening on screen.” (JAU11SP)

(68) “I never use them because if, I would read the whole time.” (FEV07SP)

However, even if many participants do not prefer to use subtitles, most participants indicate that they use subtitles under specific circumstances similar to those at T2 as can be seen in Figure 9.17.

Movie related parameters	Other languages	Language learning	Surroundings
• pace of speech • clarity of speech • accent used • type of video material	• no information about proficiency in these languages	• knowledge about word form • vocabulary knowledge	• social surrounding • local surrounding

Figure 9.17: Reasons for the use of subtitles_EE questionnaire_T3

Convergent to the findings at T2 most reasons given for the use of subtitles are connected to movie parameters, such as pace of speech (example 69), clarity of speech (example 70) or the accent of the actors (example 71), with preferences of accents again varying considerably.

(69) “I use subtitles when the speaker either has an accent or speaks too fast/mumbles.” (AMV02R)

(70) “50% depends on how good I can understand the words that are spoken. [...]” (KAG10VO)

(71) “British or Australian accent I watch with subtitles. American English usually without. (THR03VO)

Additionally, participants state that they use subtitles only for particular types of video material (examples 72-77)

(72) “I'd say 50-50. English subtitles when watching anime, no subtitles if I watch Youtube videos or movies.” (FAM02B)

(73) “I use English subtitles only five percent of the time (only with series with English as their original version). [...]” (JOD05B)

(74) “I use English subtitles most of the time when I watch English shows/movies; but not when I watch YouTube videos [...]” (JAE11CK)

(75) “0% German subtitles; 30% English subtitles (always needed for peaky Blinders because Arthur does not open his mouth while speaking).” (PAS01R)

(76) "I rarely use subtitles, but when I do I watch in English and have German subtitles (only some series like "the good doctor" where many specific words are used)". (JUM06B)

(77) "Five percent of the movies, series, I watch on Netflix are with subtitles [...]." (TIA03R)

While examples (72), (73) and (74) have a relatively broad scope, describing that JOD05B uses subtitles mainly when watching original English series, FAM02B using them when watching animes and JAE11CK using them for shows or movies but not for YouTube videos, examples (75) and (76) are more specific, referring to the use of subtitles for one particular series. Additionally, as visible in example (77) participants imply that their use of subtitles is dependent on the provider, since streaming platforms such as Netflix have integrated subtitles. Hence, the importance of movie related parameters for the use of subtitles has stayed the same from T2 to T3, even if more detailed insights are granted in the second round of data collection regarding the influence of type of video material.

Divergently, information on the use of subtitles for video content in other languages than German or English has become less rich. Many participants report to use subtitles for content in other languages but no exact indication regarding the level of proficiency in these languages has been made this time (examples 78-79).

(78) "I use English subtitles 90 of the time when I watch something that is not in German. I occasionally watch Spanish or Italian series and then I also always use English subtitles." (SOK06B)

(79) "I only use subtitles for content that is not in English (or in another language I understand well enough). If I do use them (about 30% of the time) they are always in English." (YVI10SP)

Concerning statements related to the purposeful use of subtitles for language learning a new dimension has been added in addition to acquiring knowledge about word form. While participants still appear to use subtitles consciously to enlarge their knowledge about spelling (example 80), they additionally state to use subtitles to work on their vocabulary knowledge (example 81-83).

(80) " I prefer using subtitles (so whenever I can put them on) because I instantly know how a word is spelled [...]." (NAS03R)

(81) "I use English subtitles 100% of the time. Otherwise I tend to miss new vocabulary." (LEA01SP)

(82) "Two percent most often when I am watching a movie for the second time and want to see if I understood all words correctly." (BEB05VO)

(83) "I rarely use subtitles, but when I do I watch in English and have German subtitles (only some series like "the good doctor" where many specific words are used)". (JUM06B)

As example (83) shows participants sometimes even appear to target technical words related to a particular genre. On the basis of these new findings one might claim that conscious awareness regarding new vocabulary items during video consumption has increased from T2 to T3.

Similar to the results at T2 participants also name their social and local surroundings as potential reasons for the use of subtitles. As examples (84) and (85) show participants tend to rely on subtitles when watching with less proficient people or when background noise becomes too loud (see examples 86-87).

(84) “20% (when watching with girlfriend) [...]”(THK05SP)

(85) “I only use subtitles if I watch movie in English with friends who ask to put them on [...]” (SAR07R)

(86) “Only if my environment is too loud to understand everthing. [...]” (TUS12R)

(87) “ [...] I only use them if I'm eating something and can't understand what is being said.” (CAB01R)

Similar to the results at T2 predominantly receptive EE activities were practiced at least often by over 50% of the participants. Nevertheless, the use of English apps on the phone (67%) is the most popular mixed activity, requiring both receptive and productive skills, followed by other activities (60%) and the use of search engines in English (58%). As visible in Figure 9.18 below participants’ EE activities subsumed under other activities have become even more diverse in comparison to T2.



Figure 9.18: Word cloud for other EE activities named in EE questionnaire_T3

While fewer people appear to agree on the EE tasks dreaming or writing poems, with only 1 person mentioning each of these activities, many new activities have been added to the list. For instance, one participant visits English comedy shows regularly, another student participates in English workouts and others play board games in English. Overall, only tutoring lessons have remained to be popular with three people still giving tuition in English.

Regarding solely productive actions thinking or talking to oneself is still the most popular activity with 59% doing this often at T3. This is followed by singing English songs (58%) and writing lists or notes in English (52%). Hence, no serious changes from T2 to T3 can be detected regarding the popularity of productive EE activities.

However, participants appeared to have spent slightly less time on EE activities with students being on average involved in nine instead of ten EE activities nearly always. As for the results at T2 differences between the individual students are relatively large ($IQR=9.0$, $SD=7.15$) ranging between 0 and 34 activities performed nearly always in English. However, only one person does not participate in at least one activity nearly always in English, meaning that 99% of all participants engage in at least one EE activity nearly always.

9.8 Differences in EE behaviour regarding biographical background variables

Similar to T2 the relationship between EE activities and gender, age, age of onset, highest educational degree, experience abroad and speaking opportunity will also be investigated in this order for T3.⁸⁴ Overall, data of 110 EE_questionnaires was used, revealing that some between-group differences were persistent over the time span of half a year and that other trends appear to have decreased resulting in no significant findings anymore. In the following persistent differences are going to be mentioned, since it can be assumed that tendencies which have remained unchanged over half a year are the most prominent.

9.8.1 Gender

Concerning the biographical variable of gender, differences were stable for three out of four computational variables as can be seen in Figure 9.19.

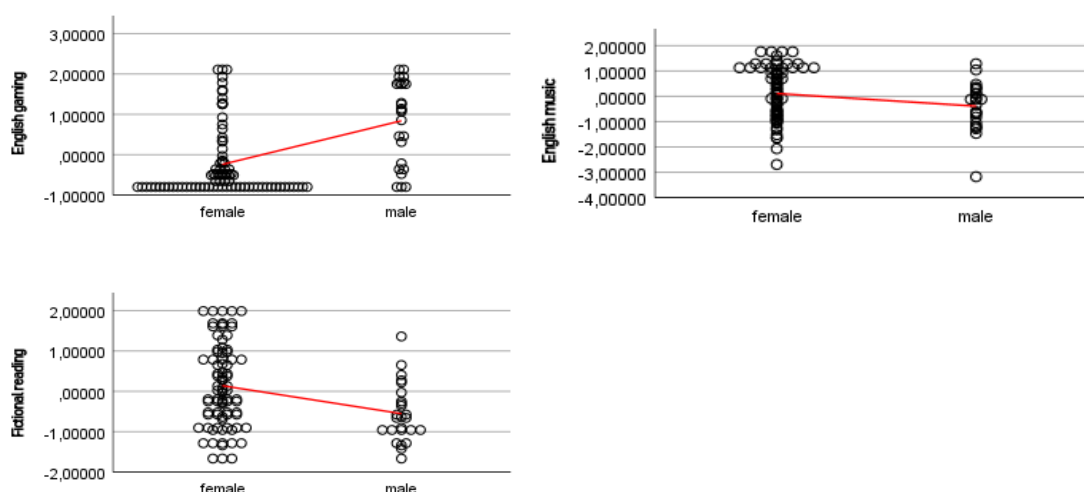


Figure 9.19: Scatterplots with interpolation line for computational variables_gender

⁸⁴ Findings on the variables of age and number of languages spoken can be found in appendix 9.1.6.

Significant differences similar to T2 were found for *English gaming* ($U=1616.5$, $z=4.494$, $p<0.001$, $r=0.437$), with men participating in all four gaming related EE activities more frequently than women. Additionally, contrasts found between male and female participants regarding *English music* ($U=680.0$, $z=-2.038$, $p=0.042$, $r=-0.200$) and *Fiction-oriented reading* ($U=585.5$, $z=-3.295$, $p=0.001$, $r=-0.319$) showed to be persistent over the time span of half a year. Regarding *English music* females outperformed their male counterparts concerning private consumption of music ($U=782.0$, $z=-1.979$, $p=0.048$, $r=-0.193$), concerts attended ($U=735.5$, $z=-2.016$, $p=0.044$, $r=-0.211$) and music videos watched ($U=704.5$, $z=-2.167$, $p=0.030$, $r=-0.197$). Similar findings were detected for reading English books ($U=1580.5$, $z=-4.629$, $p<0.001$, $r=-0.447$).

Hence, men participated significantly more frequently in diverse activities related to gaming, while women listened more to English music in various settings and read more English books over half a year. The longitudinal nature of these differences suggests that these contrasts might be more deeply engrained gender behaviours.

9.8.2 Age of onset

While all three computational variables, *English gaming* ($H(2)=7.695$, $p=0.021$, $r=-0.269$), *Written communication* ($H(2)=8.883$, $p=0.012$, $r=-0.189$) and *English speaking* ($H(2)=10.185$, $p=0.006$, $r=-0.190$), are again influenced by age of onset at T3 (see Figure 9.20), slight alterations can be located when investigating the level of individual activities.

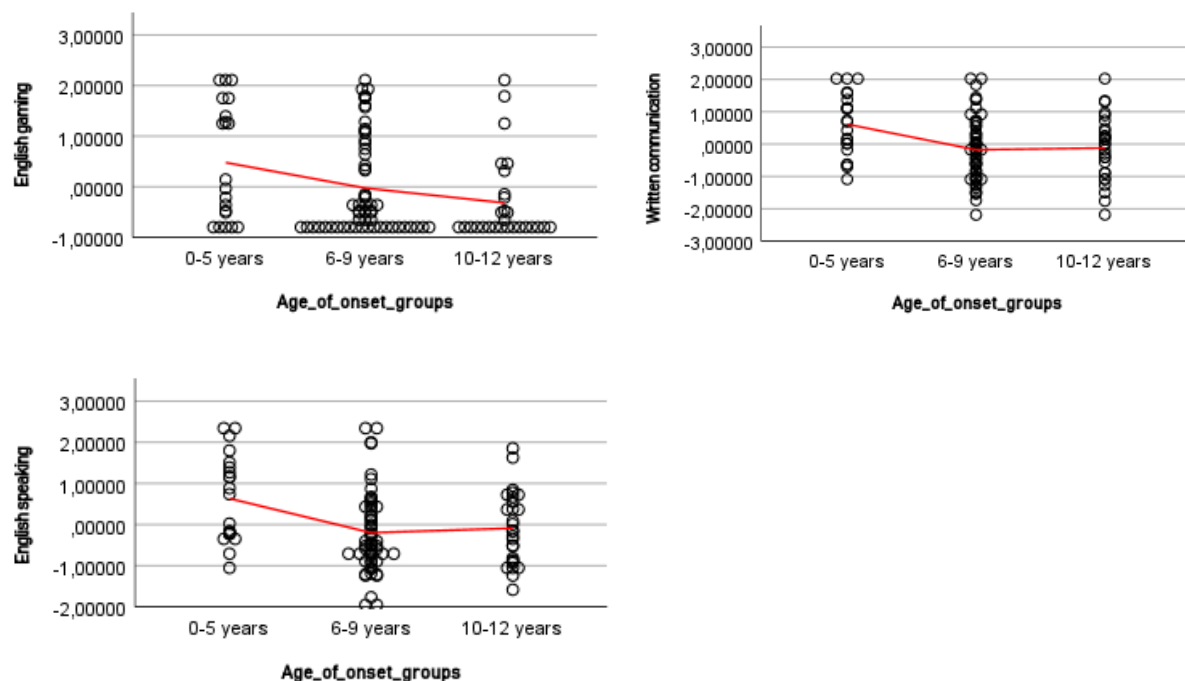


Figure 9.20: Scatterplots with interpolation line for computational variables_age of onset

Regarding *English gaming*, all participants starting to learn the English language between the age of 0-5 still engage more in playing computer games online ($H(2)=6.276$, $p=0.043$)

and offline ($H(2)=10.407$, $p=0.005$) than participants who had acquired the language in secondary school. Nevertheless, no significant differences could be found for the use of In-game chats ($H(2)=4.605$, $p=0.100$) or online communication tools ($H(2)=0.738$, $p=0.692$) anymore. Moreover, no difference between participants who were aged 6-9 and 10-12 when learning English could be detected anymore, indicating that smaller differences in age of onset can be compensated over time regarding the involvement in *English gaming*. In contrast, involvement in *Written communication* and *English speaking* only differs between the age of onset groups 0-5 and 6-9 anymore.

Hence, overall, age of onset remains to be an influential factor on *English gaming*, *Written communication* and *English speaking*. Although differences became more specific, with *English gaming* involvement differing only between the two extreme poles and *Written communication* and *English speaking* varying only between participants aged 0-5 and 6-9 when learning the English language.

9.8.3 Highest educational degree

Taking a look at the persistent differences found for participants with differing educational backgrounds reveals that educational background still influences all three computational variables, namely *Communication oriented reading* ($H(3)=9.182$, $p=0.027$, $r=-0.010$), *Written communication* ($H(3)=11.549$, $p=0.009$, $r=0.154$) and *English speaking* ($H(3)=15.323$, $p=0.002$, $r=0.131$). However, similar to the independent variable of age of onset, the effects have been partly limited to specific target groups.

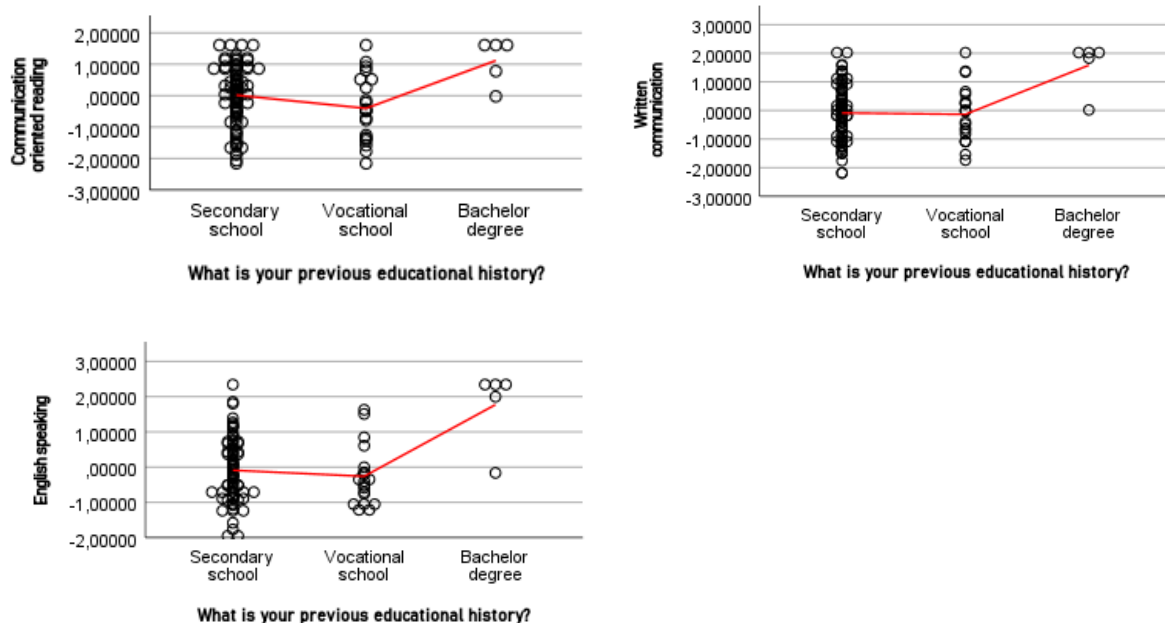


Figure 9.21: Scatterplots with interpolation line for computational variables_highest educational degree

For *Written communication* significant findings can still be found for both types of school graduates, secondary school graduates ($H(3)=-2.984$, $p=0.021$) and vocational school graduates ($H(3)=-2.921$, $p=0.021$) in comparison to participants with a BA degree.

A similar trend has been found for the engagement in *English speaking* activities. Once more a significant difference is visible between both secondary school graduates ($H(3)=-2.969$, $p=0.018$) and vocational school graduates ($H(3)=-3.209$, $p=0.008$) and participants holding a BA degree. This is partially reflected in the individual speaking activities by both types of school graduates still lagging behind regarding communication on the phone ($H(3)=-3.022$, $p=0.015$; $H(3)=-2.958$, $p=0.019$). However, face to face interactions ($H(3)=13.560$, $p=0.004$) are only done significantly less by vocational school graduates.

Divergently, no difference between participants holding a BA degree and secondary school students can be found anymore for *Communication oriented reading* ($H(3)=-2.279$, $p=0.136$) at T3. However, vocational school graduates still differ significantly from participants with a first university degree regarding their overall involvement in *Communication oriented reading* ($H(3)=-2.939$, $p=0.020$). In more detail the most persistent difference between these two groups can be located for reading English SMS or WhatsApp messages, with vocational school graduates doing this significantly less often ($H(3)=-2.612$, $p=0.054$).

Overall, more persistent differences regarding *Communication oriented reading* and *English speaking* can be found between vocational school graduates and participants holding a BA degree. However, both types of school graduates are still less involved in *Written communication* in English than participants with a first university degree.

9.8.4 Experience abroad

While experience abroad had a significant influence on involvement in English *Written communication* and *English speaking* at T2, it appears to have no impact on both variables anymore at T3 ($U=2515.0$, $z=1.664$, $p=0.096$; $U=1273.0$, $z=1.184$, $p=0.236$).

In contrast, one persistent effect could be detected for speaking more English via Skype ($U=1431.0$, $z=2.281$, $p=0.023$), with participants who had stayed abroad in an English speaking environment. However, overall experience abroad appears to have no long lasting effect on involvement on EE behaviour.

9.8.5 Speaking opportunity

The independent variable speaking opportunity, which considers whether participants regularly have a chance to speak English in their free time, appears to be a very persistent influencing factor. All differences detected at T2 were still present at T3, with significant differences found for the computational variables *English online* ($U=1891.5$, $z=3.382$, $p=0.001$, $r=0.325$), *Communication oriented reading* ($U=1962.0$, $z=3.775$, $p<0.001$, $r=0.363$), *Creative writing* ($U=2057.0$, $z=4.683$, $p<0.001$, $r=0.453$), *Written communication* ($U=2131.0$, $z=4.843$, $p<0.001$, $r=0.466$) and *English speaking* ($U=2344.5$, $z=6.391$, $p<0.001$, $r=0.618$).

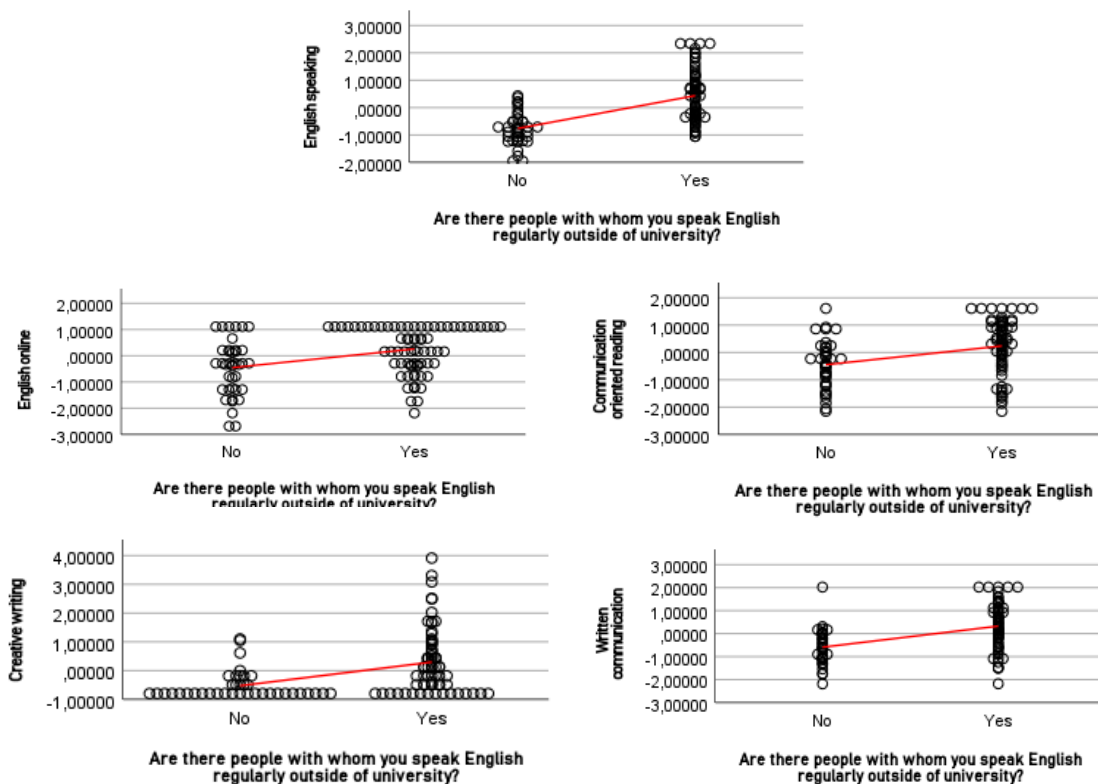


Figure 9.22: Scatterplots with interpolation line for computational variables_speaking opportunity

Overall, participants having an opportunity to speak English regularly in their spare time have used more English apps on the phone ($U=1813.5$, $z=2.979$, $p=0.003$) and tended to operate search engines more often in English ($U=1843.5$, $z=3.153$, $p=0.002$) over half a year. Moreover, they have read and written more e-mails ($U=1763.5$, $z=2.590$, $p=0.010$; $U=1848.0$, $z=3.165$, $p=0.002$), SMS or WhatsApp messages ($U=1960.0$, $z=3.885$, $p<0.001$; $U=1967.0$, $z=3.951$, $p<0.001$), social media messages ($U=1705.0$, $z=2.256$, $p=0.024$; $U=1982.0$, $z=4.013$, $p<0.001$) and status messages or comments ($U=1712.0$, $z=2.297$, $p=0.022$; $U=1956.5$, $z=3.844$, $p<0.001$). Apart from text types related to *Written communication* these participants are also more involved in writing creative texts, such as English stories ($U=1892.5$, $z=3.606$, $p<0.001$) and comics or mangas ($U=1521.0$, $z=2.182$, $p=0.029$) at both T2 and T3. Finally, the considerably higher involvement of participants stating that they have regular speaking opportunities in speaking English via Skype ($U=2126.5$, $z=5.032$, $p<0.001$), on the phone ($U=2297.5$, $z=6.318$, $p<0.001$) and in face to face interactions ($U=2051.0$, $z=4.540$, $p<0.001$) at both T2 and T3, validates students' reported EE behaviour.

Considering this impressive list of EE activities enhanced by regular spare time contact with spoken English, one can conclude that having speaking opportunities outside of class heavily influences participants' EE behaviour in a positive way.

9.9 Changes of attention to new words

No significant change is visible in the amount of attention participants pay to new words in their spare time ($T=844.0$, $z=0.929$, $p=0.353$; $T=1106.0$, $z=0.003$, $p=0.997$; $T=885.0$, $z=-0.235$, $p=0.814$, $T=376.0$, $z=0.084$, $p=0.933$). Similar to T2 participants report that they invest conscious attention to new words, which are important for the content ($Mdn=5.0$; $IQR=1.0$) and that they regularly consider new words, which they encounter in English media ($Mdn=4.0$, $IQR=1.0$). This is done significantly more often by participants who were abroad in an English speaking country ($U=1333.5$, $z=2.038$, $p=0.042$). Conversely, they state that they seldom recognize new words but do not pay conscious attention to them ($Mdn=2.0$, $IQR=1.0$) or that they are not interested in new words in their free time ($Mdn=1.0$, $IQR=1.0$).

9.10 Changes of recording techniques

Concerning the type of recording techniques selected by participants only one significant change was detected when comparing the results from T2 to the ones from T3, namely a decrease in the use of flashcards ($T=502.0$, $z=-2.141$, $p=0.032$). Hence, participant behaviour regarding recording techniques in the spare time appears to have remained unchanged, with a preference of word lists ($Mdn=3.0$; $IQR=3.0$), oral repetition ($Mdn=2.0$, $IQR=1.0$) and unorganized notes ($Mdn=2.0$; $IQR=1.0$).

9.11 Changes of word knowledge aspects

A similar picture as for the development of recording techniques can be found for the changes regarding consideration of word knowledge aspects from T2 to T3. Apart from a decline of translations ($T=353.0$, $z=-2.314$, $p=0.021$) no significant differences were found between T2 and T3.

Additionally, most trends dependent on biographical background variables have shown to be not stable over time. The only difference, which remained significant over half a year was that females consider significantly more English synonyms of a new word in their spare time than male participants ($U=713.0$, $z=-2.183$, $p=0.029$).

To summarize overall EE behaviour of participants remained relatively unchanged over the course of a semester. However, not all differences in EE behaviour regarding biographical background variables showed to be persistent. Concerning the variable of gender, observed effects only stayed stable for involvement in *English gaming*, *English music* and *Fiction-oriented reading*. Similarly participants with different age of onset consistently differed regarding engagement in *English gaming*, *Written communication* and *English speaking*. The same accounts for participants with different educational degrees, who show different behaviour concerning *Communication-oriented reading*, *Written communication* and *English speaking* over the whole time of investigation. Last but not least, the existence of the opportunity to speak English regularly in their spare time showed to have a persistent effect

on five computational variables, namely *English speaking*, *English online*, *Communication-oriented reading*, *Creative writing* and *Written communication*. Not many significant changes were observed regarding the level of attention that students paid to new words in their spare time, as well as their recording techniques and word knowledge aspects considered.

10 Results on the relation between the main concepts

While prior chapters have dealt with results concerning only one of the two researched concepts, namely vocabulary knowledge and Extramural English behaviour, this chapter will focus on potential relationships between them. Chapter 10.1 will investigate whether extramural English behaviour of the participants has any impact on their knowledge of academic verbs. Chapter 10.2 will examine whether students' biographical background has any influence on the vocabulary knowledge tested. Moreover, a linear multiple regression model investigating the influence of extramural engagement on different word knowledge aspects additional to biographical variables was calculated.

10.1 Impact of EE behaviour on vocabulary knowledge of academic verbs

In the following the impact of Extramural English behaviour on students' vocabulary knowledge of academic verbs will be examined and the following research questions will be answered:



Is there a relationship between EE activities and academic verb knowledge?

- Is there a relationship between EE activities and form-meaning recognition knowledge of academic verbs?
- Is there a relationship between EE activities and form-meaning recall knowledge of academic verbs?
- Is there a relationship between EE activities and collocation recognition knowledge of academic verbs?
- Is there a relationship between EE activities and multiple meaning recognition knowledge of academic verbs?



Which EE activities can be identified as predictors of academic verb knowledge?

- Which EE activities predict form-meaning recognition knowledge of academic verbs?
- Which EE activities predict form-meaning recall knowledge of academic verbs?
- Which EE activities predict collocation recognition knowledge of academic verbs?
- Which EE activities predict multiple meaning recognition knowledge of academic verbs?

As already outlined in chapter 9.2 the concept of Extramural English has been grouped into eleven different categories of extramural behaviour, namely involvement in English music, English video material, English online, English gaming, Fiction-oriented reading, Information oriented reading, Communication oriented reading, Written communication, Creative writing and English speaking, on basis of an exploratory factor analysis. With the help of linear multiple regression modelling it will be examined whether one or more of these eleven categories of EE behaviour have an influence on students' form-meaning recognition and form-meaning recall knowledge, as well as on their collocation recognition knowledge and their multiple meaning recognition knowledge. Moreover, it will not only be possible to detect potential relationships between EE involvement and academic vocabulary performance but it will additionally be investigated whether potential links are stable over different points in time and consequently different stages of lexical development.

10.1.1 Second data collection

As outlined in chapter 6, the first data collection point at which both data on EE and vocabulary knowledge was collected was the second data collection point. This comprises the two vocabulary tests based on form-meaning link knowledge as well as the two tests investigating multiple meaning and collocation recognition knowledge.

Based on a linear multiple regression, which is used to “examine the relationship between a criterion variable (CV, also called a dependent variable [...]) and multiple predictor variables (PVs, also called independent variables [...])” (Jeon, 2015, p. 131), these research questions are examined. In the following the linear multiple regression, proposing a linear relationship between the score on the Form-meaning Recognition Test of Academic Verbs and involvement in the eleven EE categories, will be outlined in more detail.

10.1.1.1 Form-meaning recognition

Before starting to investigate relationships between the variables all variables were examined for potential outliers or influential cases, which could influence a linear relationship between the variables. Detected outliers were winsorized⁸⁵ (Field, 2018, p. 264). To investigate which of the ten EE variables might function as reasonable predictors for form-meaning recognition knowledge of academic verbs, scatterplots were created to spot potential linear relationships between the two variables. As visible in Figure 10.1 only two variables display a reasonable linear relationship with the dependent variable, namely Fiction-oriented reading and Creative writing. This was further examined by conducting a Spearman correlation between all potential predictor variables and the outcome. As visible

⁸⁵ To counteract potential bias due to outliers, Field (2018) introduces different methods, such as trimming (has been used for the calculation of the robust ANOVA in chapter 6) and winsorizing. Since, trimming goes hand in hand with a reduction of participant number, winsorizing offers a solution to counteract bias without reducing participant numbers. Field (2018, p. 264) defines the process as follows: “Winsorizing the data involves replacing outliers with the next highest score that is not an outlier”.

in Table 10.1 only the two variables mentioned beforehand correlated significantly with the outcome variable.

	Form-meaning recognition
English music	$r_s=-0.09, p=0.36$
English video material	$r_s=-0.31, p=0.75$
English online	$r_s=0.18, p=0.07$
English gaming	$r_s=-0.04, p=0.66$
Communication oriented reading	$r_s=0.11, p=0.28$
Information oriented reading	$r_s=0.13, p=0.17$
Fiction-oriented reading	$r_s=0.32, p<0.001^{**}$
Written communication	$r_s=0.12, p=0.21$
Creative writing	$r_s=0.26, p=0.008^{**}$
English speaking	$r_s=0.14, p=0.14$

Table 10.1: Correlation coefficients form-meaning recognition knowledge and potential EE predictors

While Fiction-oriented reading showed a medium correlation coefficient of $r_s=0.32, p<0.001, r^2=0.10$, Creative writing was also positively related [$r_s=0.26, p=0.008, r^2=0.07$]. In a next step a non-parametric partial correlation was calculated for Fiction-oriented reading and form-meaning recognition knowledge controlling for Creative writing, which revealed that also the partial correlation of Fiction-oriented reading was significant [$r_s=0.23, p=0.019, r^2=0.05$]. In contrast a partial correlation of Creative writing and form-meaning recognition knowledge controlling for Fiction-oriented reading showed that the unique variance of Creative writing with the outcome variable was not significant anymore, [$r_s=0.11, p=0.13$]. Therefore, two linear regression models were computed observing the singular influence of Fiction-oriented reading on form-meaning recognition knowledge and in the second model investigating the combined influence of Fiction-oriented reading and Creative writing on the outcome variable.

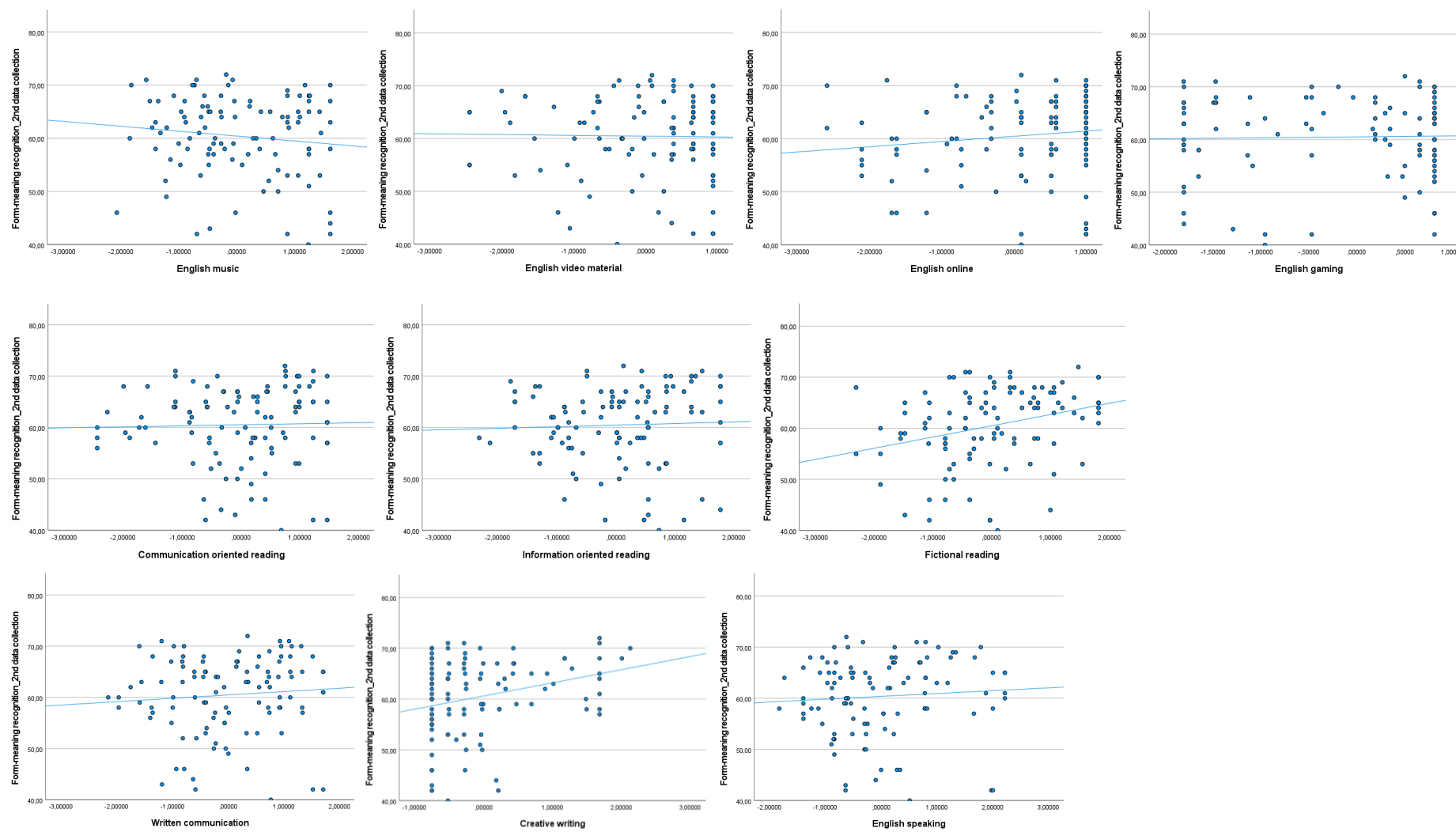


Figure 10.1: Scatterplots form-meaning recognition knowledge and potential EE predictors

As highlighted by Field (2018) in order to calculate a linear multiple regression a suitable predictor variable to sample size ratio needs to be given. Diverse recommendations (Field, Miles & Field, 2012, p. 273–274, Larson-Hall, 2016, p. 238–240) regarding the overall sample size can be found in the literature, ranging from ten participants per predictor variable (Howell, 2013, p. 527) over 15 participants (Stevens, 2009, p. 117–120) to at least 30 participants per predictor variable (Porte, 2010, p. 171). Due to the varied nature of these recommendations it has been decided to run a power analysis for linear multiple regression using Gplot (Faul et al., 2007) as suggested by Cohen et al. (2003, p. 92). To achieve a power level of 0.95 for a medium effect of 0.15 with two predictor variables a total sample size of $N=107$ is needed. If only one out of two potential predictors is used the required sample size shrinks to $N=89$. Hence $N=106$ appears to be a suitable sample size for calculating the two linear multiple regression models, as visualized in Figure 10.2 below.

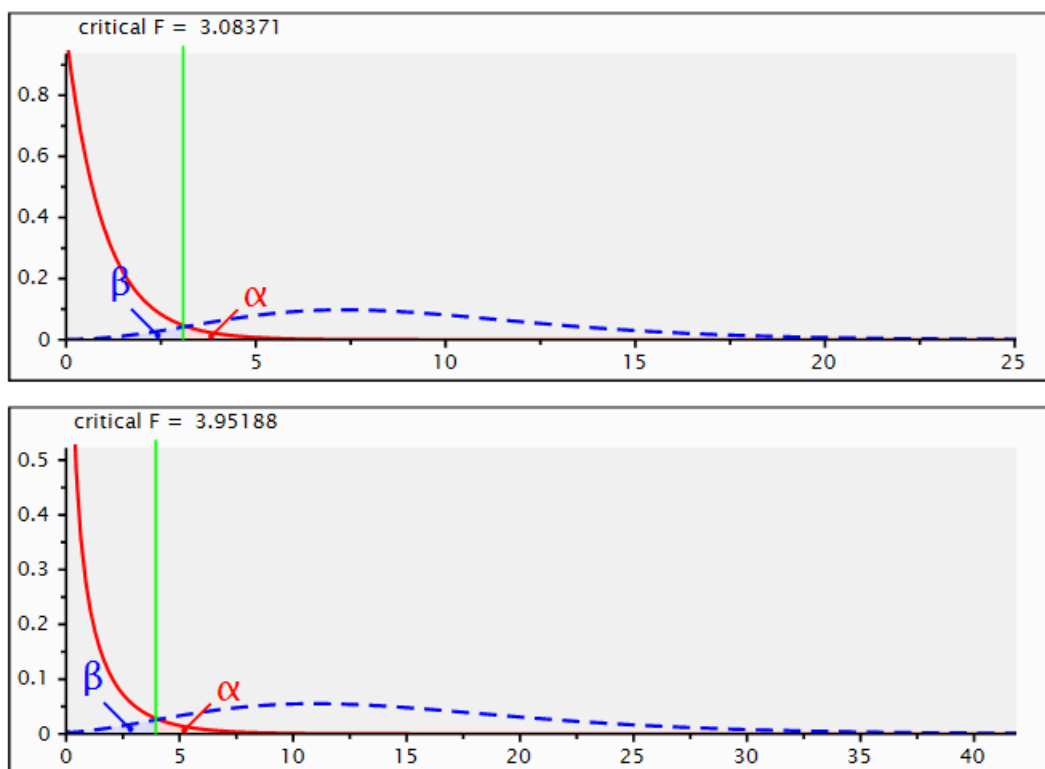


Figure 10.2: Power analysis linear multiple regression with Gplot (Faul et al., 2007)

After the evaluation of sample size the linear regression model for Fiction-oriented reading and Creative writing was computed using forced entry, which predicted the outcome significantly better than by using the mean outcome ($F(2,97)=7.831$, $p<0.001$). Before calculating the model cases 6, 39, 85, 93 and 95 were excluded due to them being outliers when looking at the standardized residuals and with them the assumption of normal distribution of errors was not given. Considering the model summary in Table 10.2 it can be stated that the two predictor variables are able to explain 13.9% of the variability of form-meaning recognition knowledge of the observed participants together.

R	R ²	R ² adjusted	Std. error of the estimate
0.373	0.139	0.121	6.112

Table 10.2: Model summary linear regression model form-meaning recognition, Fiction-oriented reading & Creative writing

In a generalized population the model would account for 12.1% of the variability. This means that if a student increases his involvement in Fiction-oriented reading and Creative writing by 1, he/she would achieve 12.1 points more on the Form-meaning Recognition Test of Academic Verbs with a maximum score of 100. Taking a look at the model parameters summarized in Table 10.3 reveals that within this model Fiction-oriented reading functions as a significant predictor of form-meaning recognition knowledge ($\beta=0.279$, $t=2.574$, $p=0.012$). In comparison Creative writing appears to be no significant predictor in combination with Fiction-oriented reading ($\beta=0.279$, $t=1.331$, $p=0.186$).

	B	95% CI	SE	T	p	β
<i>Constant</i>	62.450	[60.236, 62.664]	0.612	100.474	<0.001*	
Fiction-oriented reading	1.773	[0.406, 3.141]	0.689	2.574	0.012*	0.279
Creative writing	1.138	[-0.559, 2.835]	0.855	1.331	0.186	0.144

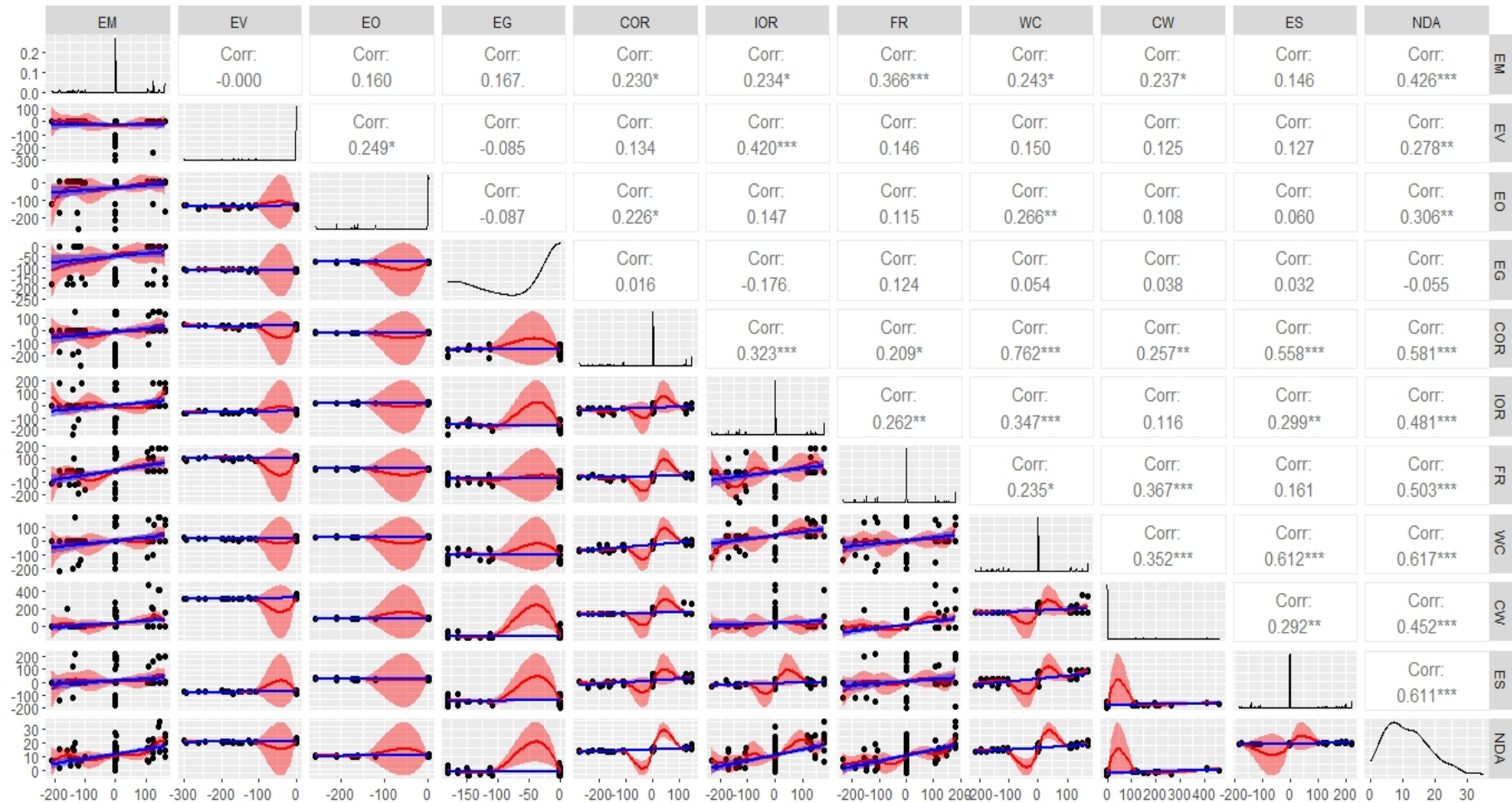
*Significant at 0.05 level

** Significant at 0.01 level

Table 10.3: Model parameters form-meaning recognition & EE predictors_second data collection

After fitting the model all assumptions for the multiple linear regression model were reviewed, which can be found in appendix 10.⁸⁶ First multicollinearity was checked by investigating the VIF and tolerance statistics, which were satisfactory. Next the assumptions of homoscedasticity and linearity were checked using scatterplots of the standardized predicted values against standardized residuals. In addition, normal distribution of unstandardized and standardized residuals was evaluated to ensure normality of errors. In a next step case-wise diagnosis was used to check for outliers. However, all standardized residuals were within the boundaries of -3 to +3 and Cook's distance, average leverage and Mahalanobis distances were within accepted limits. Finally the correlation between the two predictor variables was evaluated to ensure lack of perfect multi-collinearity (see Figure 9.3). Overall, no assumptions were violated and no bias could be detected.

⁸⁶ Similar evidence for the met assumptions for all models described in chapter 10 can be found in appendix 10 and will not be mentioned individually in each section.



EM=English music, EV=English video, EO=English online, EG=English gaming, COR=Communication oriented reading, IOR=Information oriented reading, FR=Fiction-oriented reading, WC=Written communication, CW=Creative writing, ES=English speaking, NDA=Number of daily activities

Figure 10.3: Scatterplot matrix for EE variables

Since Creative writing showed no significant individual effect on form-meaning recognition in the combined model, a linear regression model with the single predictor Fiction-oriented reading was attempted to investigate how much variability is explained by Fiction-oriented reading only. As shown by the F-test this model was significant as well, showing that the Fiction-oriented reading makes a significant improvement to predicting the outcome ($F(1,98)=13,78$, $p<0.001$). Taking a look at the model summary in Table 10.4 reveals that Fiction-oriented reading alone accounts for 12.3% of the variability in form-meaning recognition knowledge in the data set observed.

R	R ²	R ² adjusted	Std. error of the estimate
0.351	0.123	0.114	6.134

Table 10.4: Model summary linear regression model form-meaning recognition & Fiction-oriented reading

The adjusted R² indicates how well the model generalizes with a value of 11.4%, which shows that if the model were derived from the general population rather than from the observed sample, it would account for approximately 11.4% of the variance in the outcome. Again all assumptions of the linear regression model were reviewed and can be found in appendix 10.1 together with the model parameters.

In addition, computing a linear regression model with Creative writing as a single predictor reveals that it also makes a significant contribution to predicting the outcome ($F(1,98)=8.55$, $p<0.004$). As the model summary in Table 10.5 shows Creative writing accounts for 8% of the variability in form-meaning recognition knowledge in the participant sample examined and for 7% in a general population (see appendix 10.1.1).

R	R ²	R ² adjusted	Std. error of the estimate
0.283	0.080	0.071	6.285

Table 10.5: Model parameters form-meaning recognition & Fiction-oriented reading_2nd data collection

Hence, both predictor variables appear to make a significant contribution to predicting form-meaning recognition knowledge individually. However, when Fiction-oriented reading and Creative writing are combined Creative writing appears to lose parts of its explanatory power due to the relatively small unique variance explained by Creative writing as already implied by the insignificant partial correlation. When having a look at Figure 10.4 below we see that the orange and the yellow rectangles together stand for Fiction-oriented reading which explains a variability of 12.3% out of the whole surface standing for 13.9%. In contrast, Creative writing alone is able to explain 8% of the variability which is the red and the orange area together. When taken in combination, the two predictors have a relatively large overlap of 6.4%, which is the orange area. This orange area could be explained by either Fiction-oriented reading or Creative writing. Moreover, it is visible that the variability

explained by Fiction-oriented reading alone is 5.9%, which is considerably larger than what can be explained by Creative writing alone.

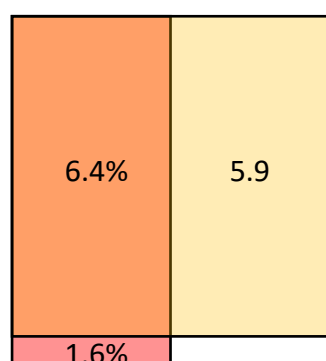


Figure 10.4: Shared & unique explained variability of form-meaning recognition by Fiction-oriented reading & Creative writing_2nd collection

Therefore, if a student asks how he or she can improve their form-meaning recognition knowledge of academic verbs by engaging in extramural English activities, we can advise him/her to involve in Fiction-oriented reading, if the maximum improvement with the minimal effort should be achieved. However, if the student wants to improve his/her form-meaning recognition knowledge maximally by engaging in EE activities, we can tell him/her to engage in both Fiction-oriented reading and Creative writing.

10.1.1.2 Form-meaning recall

In contrast to form-meaning recognition knowledge no significant Spearman or Kendal tau correlations could be found between the ten potential EE predictor variables and form-meaning recall knowledge after winsorizing.

	Form-meaning recognition
English music	$r_s = -0.16, p = 0.106$
English video material	$r_s = -0.16, p = 0.102$
English online	$r_s = 0.13, p = 0.183$
English gaming	$r_s = -0.05, p = 0.629$
Communication oriented reading	$r_s = 0.12, p = 0.218$
Information oriented reading	$r_s = 0.04, p = 0.706$
Fiction-oriented reading	$r_s = 0.18, p = 0.075$
Written communication	$r_s = 0.12, p = 0.205$
Creative writing	$r_s = 0.16, p = 0.099$
English speaking	$r_s = 0.06, p = 0.537$

Table 10.6: Correlation coefficients form-meaning recall and potential EE predictors

However, when analysing the scatterplots visible in appendix 10.1.2 in more detail it seems that again Creative writing and Fiction-oriented reading have the strongest linear relationship to form-meaning recall knowledge, which is why a linear regression model was attempted for these predictors. Since the sample size $N=106$ stays the same for all word knowledge components researched the necessary number of cases is given. Taking a look at the linear regression model including the predictor Fiction-oriented reading reveals that the

model makes no significant contribution to predicting the outcome ($F(1,103)=3.070$, $p=0.083$). Therefore, students' involvement in Fiction-oriented reading appears to have no influence on their form-meaning recall knowledge. In contrast, the linear regression model for the predictor Creative writing resulted in a significant improvement of outcome prediction through the proposed model ($F(1,104)=4.214$, $p=0.043$). Table 10.7 and Table 10.8 reveal though, that engagement in Creative writing accounts for only 3.9% of the variability in form-meaning recall of the participants in the collected sample and for 3% in a generalized population.

R	R ²	R ² adjusted	Std. error of the estimate
0.373	0.039	0.003	6.112

Table 10.7: Model summary linear regression model form-meaning recall & Creative writing

	B	95% CI	SE	T	p	β
<i>Constant</i>	42.188	[40.379, 43.997]	0.912	46.246	<0.001*	
Creative writing	2.310	[-0.079, 4.542]	1.126	2.053	0.043*	0.197

Table 10.8: Model parameters form-meaning recall & Creative writing_2nd data collection

This means that if a student increases their involvement in Creative writing by 1, they would achieve 3 points more on a form-meaning recall test on academic verb knowledge with a maximum score of 100.

Taking a look at the assumptions for the model displayed in appendix 10.1.2 reveals that all preconditions for interpreting the model are fulfilled apart from normality of residuals. Hence, in a next step outliers 6, 10, 24, 36, 61, 68, 69, 76, 83, 93 and 95, which were identified via the casewise diagnostics were excluded. After recalculating the linear regression model without these outliers the assumption of normality of residuals is given ($D(95)=0.081$, $p=0.146$) and the explanatory power of the model has increased from 3.9% to 5.5% in the participant group examined ($F(1,93)=5.441$, $p=0.022$). The same accounts for a generalized sample in which involvement in Creative writing now accounts for at least 4.5% instead of 3%.

Therefore, while Creative writing played a rather minor role as a predictor for form-meaning recognition knowledge of academic verbs in comparison to Fiction-oriented reading, Creative writing is a significant predictor for form-meaning recall knowledge in contrast to Fiction-oriented reading. Hence, students who want to improve their form-meaning recall knowledge of academic verbs through extramural involvement in an English activity, should focus on writing creative texts.

10.1.1.3 Collocation recognition

Analysing the Spearman correlation matrix for collocation recognition knowledge and the ten potential predictors of EE reveals that only two variables show a significant linear relationship with the dependent variable, namely Creative writing ($r_s=0.23$, $p=0.021$, $r^2=0.05$) and English speaking ($r_s=0.20$, $p=0.044$, $r^2=0.04$).

	Form-meaning recognition
English music	$r_s=-0.08$, $p=0.443$
English video material	$r_s=-0.02$, $p=0.810$
English online	$r_s=0.17$, $p=0.090$
English gaming	$r_s=0.10$, $p=0.315$
Communication oriented reading	$r_s=0.13$, $p=0.201$
Information oriented reading	$r_s=0.16$, $p=0.106$
Fiction-oriented reading	$r_s=0.14$, $p=0.143$
Written communication	$r_s=0.17$, $p=0.082$
Creative writing	$r_s=0.23$, $p=0.021^*$
English speaking	$r_s=0.20$, $p=0.044^*$

Table 10.9: Correlation coefficients collocation recognition and potential EE predictors

This is additionally confirmed by analysing the scatterplots for all potential predictor variables. As above sample size adequacy is given for a multiple linear regression model containing two predictor variables.

First a linear regression model for the single predictor variable Creative writing was attempted, which already resulted in a significant improvement in prediction ($F(1,99)=7.845$, $p=0.006$). For this model five cases were excluded due to them being outliers diagnosed via the casewise diagnostics. Overall, the model revealed that Creative writing accounts for 7.3% of the variability in collocation recognition knowledge in the sample examined and would explain 6.4% in a generalized population as can be seen in Table 10.10.

R	R²	R² adjusted	Std. error of the estimate
0.271	0.073	0.064	3.206

Table 10.10: Model summary linear regression model collocation recognition & Creative writing

Secondly, a linear regression model including only the independent variable English speaking was computed displaying a significant improvement of explanatory power through the model ($F(1,99)=5.885$, $p=0.017$). As can be seen in Table 10.11 English speaking explains 5.6% of the variability in collocation recognition knowledge for the participants observed and 4.7% in a generalized sample.

R	R ²	R ² adjusted	Std. error of the estimate
0.237	0.056	0.047	3.236

Table 10.11: Model summary linear regression model collocation recognition & English speaking

Hence, both Creative writing and English speaking activities appear to be significant predictors of collocation recognition knowledge. However, when analysing a multiple linear regression model of both predictors together using forced entry reveals that while both predictors together result in a significant ($F(1,99)=5.395$, $p=0.006$) improvement of collocation knowledge of 9.9% in the given sample and 8.1% in a generalized sample, only Creative writing still has a significant impact as can be seen in Table 10.12.

	B	95% CI	SE	t	p	β
<i>Constant</i>	16.730	[16.102, 17.358]	0.316	52.865	<0.001**	
Creative writing	0.879	[0.073, 1.684]	0.406	2.165	0.033*	0.218
English speaking	0.549	[-0.102, 1.199]	0.328	1.674	0.097	0.169

Table 10.12: Model parameters collocation recognition & EE predictors_2nd data collection

This indicates, as explained in Figure 10.5, that the two variables have a relatively large shared variance of 3%, which is also inferred by a non-parametric partial correlation.

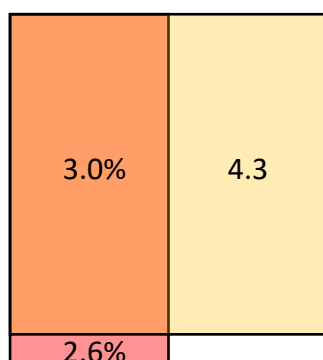


Figure 10.5: Shared & unique explained variability of collocation recognition by Creative writing and English speaking_2nd collection

Consequently, the unique variance of 2.6% of English speaking appears to be too small to be significant in combination with Creative writing, which has a unique variance of 4.3%. Therefore, students aiming at improving their collocation recognition knowledge of academic verbs by involving in only one category of EE activities will make the largest improvement when they engage in Creative writing tasks. Yet, students who would like to make the maximum improvement possible should consider to participate in both Creative writing and English speaking activities.

10.1.1.4 Multiple meaning recognition

Taking a look at the distribution of multiple meaning recognition reveals four outliers, namely cases 99, 48, 88 and 29, which were again winsorized before creating a correlation matrix between the outcome variable and the ten EE predictors. As can be seen in Table 10.13 below multiple meaning recognition knowledge appears to have only two significant linear relationships with Fiction-oriented reading and Creative writing.

	Form-meaning recognition
English music	$r_s=-0.03, p=0.767$
English video material	$r_s=0.15, p=0.136$
English online	$r_s=0.15, p=0.125$
English gaming	$r_s=0.05, p=0.581$
Communication oriented reading	$r_s=0.13, p=0.194$
Information oriented reading	$r_s=0.06, p=0.559$
Fiction-oriented reading	$r_s=0.27, p=0.005^*$
Written communication	$r_s=0.12, p=0.229$
Creative writing	$r_s=0.21, p=0.029^*$
English speaking	$r_s=0.01, p=0.931$

Table 10.13: Correlation coefficients multiple meaning recognition and potential EE predictors

This is additionally confirmed by the scatterplots, which visualize the linear relationship between multiple meaning recognition and all potential predictor variables.

Similar as for the other word knowledge components first single linear regression models were calculated for Fiction-oriented reading and Creative writing. Both models using either Fiction-oriented reading ($F(1,99)=11.450, p=0.001$) or Creative writing ($F(1,99)=5.173, p=0.025$) as predictors were significant. As visible in Table 10.14 Fiction-oriented reading alone explains a relative large amount of variability in the given sample with 10.4% and 9.5% in a generalized population.

Predictor	R	R²	R² adjusted	Std. error of the estimate
Fiction-oriented reading	0.322	0.104	0.095	1.707
Creative writing	0.223	0.050	0.040	1.757

Table 10.14: Model summary linear regression model multiple meaning recognition & EE predictors

In comparison the contribution of Creative writing to the predictability of multiple meaning recognition knowledge is only half the size. Therefore, it does not come as a surprise that in a combined multiple regression model Creative writing has no significant predictive power anymore due to its relatively small unique variance as visible in Table 10.15.

	B	95% CI	SE	t	p	β
<i>Constant</i>	19.067	[18.729, 19.405]	0.170	111.943	<0.001**	
Fiction-oriented reading	0.489	[0.108, 0.869]	0.192	2.549	0.012*	0.281
Creative writing	0.179	[-0.295, 0.653]	0.239	0.748	0.456	0.082

Table 10.15: Model parameters multiple meaning recognition & EE predictors_2nd collection

Nonetheless, the predictive power of the model increases from 10.4% in the single linear regression model to 10.9% in the multiple linear regression model in the given sample but even decreases from 9.5% to 9.1% in a generalized population. Hence, while Creative writing appears to make a small valuable addition for the students participating in this study, in a larger population Fiction-oriented reading alone would result in a larger improvement of multiple meaning recognition knowledge. Therefore, students seeking advice concerning the type of EE activity, which enhances multiple meaning recognition knowledge the most should be guided to involve in Fiction-oriented reading activities and in Creative writing activities.

Furthermore, it was investigated whether knowledge of the primary or secondary meaning of academic verbs is influenced by different EE categories.⁸⁷ Taking a first look at the correlation matrix for knowledge of primary meaning with all EE categories reveals similar findings as for overall multiple meaning recognition knowledge, with Fiction-oriented reading ($r_s=0.29$, $p=0.002$, $r^2=0.09$) and Creative writing ($r_s=0.23$, $p=0.020$, $r^2=0.05$) being positively linearly related. In comparison, knowledge of the secondary meaning of academic verbs only shows a positive linear relationship to English online ($r_s=0.21$, $p=0.033$, $r^2=0.04$).

Calculating single linear regression models for Fiction-oriented reading and Creative writing ($F(1,104)=7.191$, $p=0.098$) as predictor variables for primary meaning knowledge of academic verbs reveals that only Fiction-oriented reading appears to make a significant contribution to predicting knowledge of primary meaning ($F(1,103)=22.001$, $p=0.003$) as can also be seen in Table 10.16.

	B	95% CI	SE	t	p	β
<i>Constant</i>	9.156	[8.855, 9.458]	0.152	60.274	<0.001*	
Fiction-oriented reading	0.454	[0.155, 0.753]	0.151	3.013	0.003*	0.285

Table 10.16: Model parameters multiple meaning recognition & Fiction-oriented reading_2nd data collection

⁸⁷ Primary and secondary meaning were distinguished by checking which of the multiple meanings was mentioned first in the dictionaries used for consultation.

Nevertheless, the contribution of Fiction-oriented reading to explaining the variability of primary meaning knowledge is smaller with 8.1% in the given and 7.2% in a generalized population, than for overall multiple meaning recognition knowledge.

Calculating a linear regression model for knowledge of secondary meaning of academic verbs revealed that English online appears to make a significant improvement to predicting the outcome variable ($F(1,104)=5.473$, $p=0.047$). However, bootstrapping was used to generate the confidence intervals and significance values, since the normality of errors was not given for the model (see Table 10.17).

	B	95% CI	SE	t	p	β
<i>Constant</i>	9.928	[9.703, 10.152]	0.113	87.727	<0.001*	
English online	0.222	[0.003, 0.441]	0.193	2.009	0.047*	0.193

Table 10.17: Model parameters multiple meaning recognition & English online_2nd data collection

This resulted in a R^2 of 3.0% for the given population and an adjusted R^2 of 2.8% in a generalized population. Therefore, English online appears to make a significant but very small contribution to improving knowledge of secondary meanings of academic verbs.

10.1.2 Third data collection

As for the second data collection all four tests were administered and analysed for the third data collection.

10.1.2.1 Form meaning recognition

In a first step the dependent variable form meaning recognition and the ten potential predictor variables were checked for outliers through the generation of boxplots. All outliers found were winsorized⁸⁸. In a next step a Spearman correlation matrix was generated to examine which EE variables display a linear relationship with form-meaning recognition. As visible in Table 10.18 Fiction-oriented reading and Creative writing are linearly related to the outcome variable similar as in the second data collection. This is visualized as well in the scatterplots in appendix 10.2.1.

	Form-meaning recognition
English music	$r_s=-0.14$, $p=0.162$
English video material	$r_s=-0.12$, $p=0.213$
English online	$r_s=0.07$, $p=0.483$
English gaming	$r_s=-0.10$, $p=0.327$
Communication oriented reading	$r_s=-0.03$, $p=0.751$

⁸⁸ To counteract potential bias due to outliers, Field (2018) introduces different methods, such as trimming (has been used for the calculation of the robust ANOVA in chapter 6) and winsorizing. Since, trimming goes hand in a hand with a reduction of participant number, winsorizing offers a solution to counteract bias without reducing participant numbers. Field (2018, p. 264) defines the process as follows: "Winsorizing the data involves replacing outliers with the next highest score that is not an outlier".

Information oriented reading	$r_s=0.00, p=0.999$
Fiction-oriented reading	$r_s=0.20, p=0.044^*$
Written communication	$r_s=-0.01, p=0.914$
Creative writing	$r_s=0.26, p=0.008^*$
English speaking	$r_s=0.09, p=0.348$

Table 10.18: Correlation coefficients form-meaning recognition and potential EE predictors_3rd data collection

However, it is already visible that in comparison to the second data collection point Creative writing displays a stronger correlation, while Fiction-oriented reading appears to have lost some ground. This is confirmed by calculating a hierarchical linear multiple regression model with two blocks entered via forced entry. When first entering only Creative writing, based on its stronger linear relationship shown in the correlation coefficient, this single linear regression model already shows to be significantly better at predicting the outcome than using the mean outcome ($F(1,93)=10.113, p=0.002$). In a next step, Fiction-oriented reading is added to the model. However, the change statistics shows that the change in R^2 is not significant anymore ($p=0.634$). This is also reflected in the amount of variability in the outcome accounted for by each predictor as can be seen in Table 10.19.

Predictor	R	R²	R² adjusted	Std. error of the estimate
Creative writing	0.313	0.098	0.081	4.944
Creative writing + Fiction-oriented reading	0.317	0.100	0.088	4.965

Table 10.19: Model summary linear regression form-meaning recognition & EE predictors_3rd data collection

While Creative writing alone accounts for 9.8% of variability in the outcome variable in the observed population, the explanatory power only increases by 0.2% when adding Fiction-oriented reading. A similar small growth can be observed for the variability accounted for in a generalized sample.

Moreover, an analysis of the model parameters in Table 10.20 further supports this conclusion by revealing that Fiction-oriented reading does not make a significant contribution to predicting the outcome in the combined model.

	B	95% CI	SE	t	p	B
<i>Constant</i>	64.059	[63.046, 65,072]	0.510	125.612	<0.001*	
Creative writing	1.851	[0.422, 3.279]	0.719	2.573	0.012*	0.288
Fiction-oriented reading	0.271	[-0.857, 1.400]	0.568	0.477	0.634	0.053

Table 10.20: Model parameters form-meaning recognition & EE predictors_3rd data collection

In comparison to the second data collection displaying a strong predictive power of Fiction-oriented reading, this predictor variable has been replaced by Creative writing, which was still a minor significant predictor in the first semester observed.

10.1.2.2 Form meaning recall

While no significant correlations could be found after one semester of studies between form meaning recall knowledge of academic verbs and the ten potential EE predictor variables, two significant linear relationships could be detected using Spearman correlations after a second semester of studying as summarized in Table 10.21.

	Form-meaning recall
English music	$r_s = -0.11, p = 0.286$
English video material	$r_s = -0.13, p = 0.173$
English online	$r_s = 0.07, p = 0.506$
English gaming	$r_s = -0.08, p = 0.418$
Communication oriented reading	$r_s = -0.04, p = 0.659$
Information oriented reading	$r_s = 0.07, p = 0.442$
Fiction-oriented reading	$r_s = 0.20, p = 0.041^*$
Written communication	$r_s = 0.09, p = 0.362$
Creative writing	$r_s = 0.26, p = 0.008^*$
English speaking	$r_s = 0.13, p = 0.194$

Table 10.21: Correlation coefficients form-meaning recall and potential EE predictors_3rd data collection

A similar picture is produced in the scatterplots of form-meaning recall knowledge and the ten EE variables, showing that only Fiction-oriented reading and Creative writing display a stronger linear relationship to the outcome variable.

Computing a single linear regression model with the predictor of Fiction-oriented reading reveals that involvement in Fiction-oriented reading significantly improves the prediction of the outcome variable ($F(1,98) = 7.860, p = 0.006$) after excluding five outliers of standardized residuals (see Table 10.22).

	B	95% CI	SE	t	p	B
<i>Constant</i>	46.375	[44.802, 47.949]	0.793	58.492	<0.001*	
Fiction-oriented reading	2.213	[0.646, 2.779]	0.789	2.804	0.006*	0.272

Table 10.22: Model parameters form-meaning recall & Fiction-oriented reading_3rd data collection

Moreover, Fiction-oriented reading as a single predictor of form-meaning recall knowledge of academic verbs is able to explain 7.4% of the variability in the observed sample and 6.5% in a generalized population.

Examining the explanatory power of creative writing as a single predictor, reveals that similar to Fiction-oriented reading Creative writing makes a significant improvement to the prediction of form-meaning recall knowledge of academic verbs ($F(1, 98)=13.947$, $p<0.001$) as also visible in Table 10.23.

	B	95% CI	SE	t	p	B
<i>Constant</i>	46.506	[44.973, 48.040]	0.773	60.190	<0.001**	
Creative writing	3.634	[1.703, 5.564]	0.973	3.735	<0.001**	0.353

Table 10.23: Model parameters form-meaning recall & Creative writing_3rd data collection

Comparing Creative writing as a single predictor to Fiction-oriented reading reveals that engagement in Creative writing explains even more variability of form-meaning recall knowledge, namely 12.5% in the student population examined and 11.6% in a global population.

Hence, one could assume that the two predictors together will explain a relatively large percentage of the variability of form-meaning recall knowledge. However, as a multiple linear regression model using forced entry shows that despite making a significant contribution to predicting the model ($F(2,97)=7.778$, $p<0.001$), the two EE predictors together only explain 13.8% in the observed sample and 12.0% in a general sample. Considering Table 10.24 below shows that in a multiple linear regression model Fiction-oriented reading makes no significant contribution to predicting form-meaning recall knowledge anymore.

	B	95% CI	SE	t	p	B
<i>Constant</i>	46.512	[44.982, 48.041]	0.771	60.359	<0.001**	
Fiction-oriented reading	1.082	[-0.652, 2.816]	0.874	1.238	0.219	0.113
Creative writing	2.972	[0.774, 5.171]	1.108	2.683	0.009*	0.289

Table 10.24: Model parameters form-meaning recall & EE predictors_3rd data collection

Again this phenomenon can be explained by the fact that the two predictor variables share a large percentage of variance, which leaves Fiction-oriented reading with too little unique variance explained to be significant after the shared variance with Creative writing has been extracted.

10.1.2.3 Collocation recognition

While collocation recognition knowledge was predicted significantly by English speaking and Creative writing after one semester of studies, no significant correlations can be found for these to EE variables in the second semester of studying (see Table 10.25).

	Collocation recognition
English music	$r_s = -0.08, p = 0.410$
English video material	$r_s = 0.04, p = 0.704$
English online	$r_s = 0.21, p = 0.035^*$
English gaming	$r_s = -0.03, p = 0.753$
Communication oriented reading	$r_s = 0.06, p = 0.580$
Information oriented reading	$r_s = 0.14, p = 0.147$
Fiction-oriented reading	$r_s = 0.22, p = 0.022^*$
Written communication	$r_s = 0.08, p = 0.404$
Creative writing	$r_s = 0.18, p = 0.066$
English speaking	$r_s = 0.16, p = 0.107$

Table 10.25: Correlation coefficients collocation recognition and potential EE predictors_3rd data collection

Instead two significant linear relationships were detected for English online ($r_s = 0.21, p = 0.035$) and Fiction-oriented reading ($r_s = 0.22, p = 0.022$).

Taking a similar approach as for investigating the predictors for form-meaning recall knowledge, linear regression models are calculated for each of the two variables independently first. As visible in Table 10.26 below English online makes a significant contribution to predicting collocation recognition knowledge of academic verbs in the second semester observed ($F(1,99) = 6.497, p = 0.012$).

	B	95% CI	SE	t	p	β
Constant	17.668	[17.019, 18.316]	0.327	54.082	<0.001**	
English online	0.822	[0.182, 1.461]	0.322	2.549	<0.012*	0.248

Table 10.26: Model parameters collocation recognition & English online_3rd data collection

As single predictor involvement in English online activities explain 6.2% of the variability of collocation recognition knowledge of the students taking part in the study and would generalize to 5.2%.

Interpreting the linear regression model including Fiction-oriented reading reveals a similar significant improvement in prediction of collocation recognition knowledge in the second semester observed ($F(1,99) = 6.629, p = 0.012$).

	B	95% CI	SE	t	p	β
Constant	17.722	[17.075, 18.370]	0.326	54.318	<0.001**	
Fiction-oriented reading	0.834	[0.191, 1.476]	0.324	2.575	<0.012*	0.251

Table 10.27: Model parameters collocation recognition & Fiction-oriented reading_3rd data collection

Moreover, both predictors appear to have a similar predictive power, if interpreted alone, since Fiction-oriented reading explains 6.3% of the variance in the sample observed and 5.3% in a generalized sample.

In contrast to form-meaning recall knowledge collocation recognition knowledge, seems to be predicted significantly by both English online and Fiction-oriented reading activities ($F(1,98)=5.332$, $p=0.006$), as a multiple linear regression using forced entry shows (see Table 10.28).

	B	95% CI	SE	t	p	β
<i>Constant</i>	17.690	[17.051, 18.330]	0.322	54.924	<0.001**	
English online	0.646	[-0.008, 1.300]	0.330	1.961	0.052*	0.195
Fiction-oriented reading	0.660	[0.003, 1.317]	0.331	1.993	0.049*	0.198

Table 10.28: Model parameters collocation recognition & EE predictors_3rd data collection

Together English online and Fiction-oriented reading predict 9.8% of the variability in the given sample and 8% in a generalized population. Hence, the predictors both explain enough unique variance and complement each other.

10.1.2.4 Multiple meaning recognition

Before starting to analyse potential predictors for multiple meaning recognition knowledge after another semester of studying, three outliers, namely 86, 90 and 100, had to be winsorized. In a next step calculating Spearman correlations revealed that only Fiction-oriented reading appears to have a significant linear relationship with the outcome variable as visible in Table 10.29.

	Multiple meaning recognition
English music	$r_s=-0.05$, $p=0.627$
English video material	$r_s=0.002$, $p=0.984$
English online	$r_s=0.08$, $p=0.423$
English gaming	$r_s=-0.04$, $p=0.660$
Communication oriented reading	$r_s=0.11$, $p=0.267$
Information oriented reading	$r_s=0.10$, $p=0.333$
Fiction-oriented reading	$r_s=0.28$, $p=0.004^*$
Written communication	$r_s=0.13$, $p=0.171$
Creative writing	$r_s=0.16$, $p=0.096$
English speaking	$r_s=0.19$, $p=0.058$

Table 10.29: Correlation coefficients multiple meaning recognition and potential EE predictors_3rd data collection

This means that the second predictor after one semester of studying, namely Creative writing has lost its predictive potential. This is validated by calculating a hierarchical multiple linear regression model with two blocks using forced entry. As visible in Table 10.30 Creative

writing makes no significant contribution to predicting multiple meaning knowledge, while Fiction-oriented reading alone does ($F(1, 103)=7.534, p=0.007$).

	B	95% CI	SE	t	p	β
<i>Constant</i>	19.329	[18.891, 19.767]	0.221	87.378	<0.001*	
Fiction-oriented reading	0.486	[-0.013, 0.986]	0.252	1.930	0.056	0.211
Creative writing	0.295	[-0.338, 0.928]	0.319	0.925	0.357	0.101

Table 10.30: Model parameters multiple meaning recognition & EE predictors_3rd data collection

However, as visible in Table 10.31 adding Creative writing would still add explanatory power for the sample observed by increasing it from 6.8% explained by Fiction-oriented reading only to 7.6%. In a generalized population however, it would even lower the explained variability.

Predictor	R	R²	R² adjusted	Std. error of the estimate
Fiction-oriented reading	0.261	0.068	0.059	2.260
Fiction-oriented reading + Creative writing	0.276	0.076	0.058	2.262

Table 10.31: Model summary linear regression multiple meaning recognition & EE predictors_3rd data collection

Hence, EE activities related to Fiction-oriented reading appear to be the only significant predictors of multiple meaning recognition knowledge of academic verbs after two semesters of observation.

10.1.3 Longitudinal aspect

While the above chapters have presented the results for each data collection largely independently of each other this chapter will try to connect these jigsaws to a somewhat more complete puzzle. While most prior studies investigating the influence of EE on vocabulary knowledge were not longitudinal in nature and rather relied on correlational statistical methods (Lee, 2019a; Li & Hafner, 2021; Noori, Gholami & Rajabi, 2014), this study has the benefit of being longitudinal, following the same participant group for two semesters in addition to using more complex statistical procedures. This combination enables a unique view on the development of the relationship between EE predictor variables and their influence on four different aspects of word knowledge, investigated through the following research question:

🕒 *How stable are the relationships between EE activities and the four different types of academic verb knowledge over one academic year?*

As visible in Table 10.32 below some regularities regarding the influence of EE on the different word knowledge aspects can be detected.

	Second data collection	Third data collection
Form-meaning recognition	Fiction-oriented reading Creative writing*	Creative writing Fiction-oriented reading*
Form-meaning recall	Creative writing	Creative writing Fiction-oriented reading*
Collocation recognition	Creative writing English speaking*	English online Fiction-oriented reading
Multiple meaning recognition	Fiction-oriented reading Creative writing*	Fiction-oriented reading

** EE categories with an asterisk were found to be significant predictors if performed in isolation. They lost their predictive power when being added into a multiple linear regression model with other predictors.*

Table 10.32: Summary of linear regression results for EE and all vocabulary tests

Overall, Creative writing and Fiction-oriented reading appear to be the EE activities with the largest impact on all different types of academic verb knowledge investigated. Three out of the four word knowledge components, namely form-meaning recognition, form-meaning recall and multiple meaning recognition, are only explained by one or both of these two EE activities. Only collocation recognition knowledge appears to be influenced by additional EE activities, namely English speaking and English online, being the only word knowledge component observed categorized under word use and not word meaning in Nation's (2013) taxonomy.

When recalling again which EE activities are subsumed under these factors reveals that Creative writing entails producing English stories, fanfiction, comics and blogs, while Fiction-oriented reading encompasses the receptive consumption of similar genres, such as English books, short stories and fanfiction. Hence, engaging with these genres appears to be particularly fruitful for all word knowledge aspects in regard to academic verb knowledge.

Concerning the additional EE activities, English speaking, which influence collocation recognition knowledge, English speaking includes language production via internet services,

on the phone, face to face or to yourself. Moreover, English online is composed of using English apps and search engines. While the necessity of collocation knowledge appears apparent for successful participation in English speaking activities, English online only reveals its connection at second thought. Both applications subsumed under English online require the entry of search terms, which oblige the user to think about word combinations.

This shows that in sum the most influential EE activities appear to be productive, apart from Fiction-oriented reading, which is a receptive EE activity, that requires more intensive engagement with the English language in comparison to other receptive EE activities.

Taking a look at the development of the total variance explained for the group of participants observed in Table 10.33 at both data collection points, shows that regarding explanatory power of variance there is no uniform picture for all test types.

	2 nd data collection	3 rd data collection
Form-meaning recognition	$R^2=8.0\%$, $R^2_a=7.1\%$	$R^2=9.8\%$, $R^2_a=8.8\%$
Form-meaning recall	$R^2=5.5\%$, $R^2_a=4.5\%$	$R^2=13.8\%$, $R^2_a=12.0\%$
Collocation recognition	$R^2=9.9\%$, $R^2_a=8.1\%$	$R^2=9.8\%$, $R^2_a=8.0\%$
Multiple meaning recognition	$R^2=10.4\%$, $R^2_a=9.5\%$	$R^2=6.8\%$, $R^2_a=5.9\%$

Table 10.33: Variability explained in the population observed

As also visible in Figure 10.6 depicting these developments for a generalized sample, the influence of the EE activities increases for both form-meaning recognition and form-meaning recall. However, while form-meaning recognition knowledge shows only a slight increase of 1.8%, the explanatory power is nearly tripled for form-meaning recall knowledge. Linking this to the types of EE activities found to influence form-meaning knowledge reveals that form-meaning recall knowledge is the only word knowledge component which is consistently influenced by the same EE activity, namely Creative writing. In comparison form-meaning recognition knowledge is influenced alternately by Creative writing and Fiction-oriented reading the most.

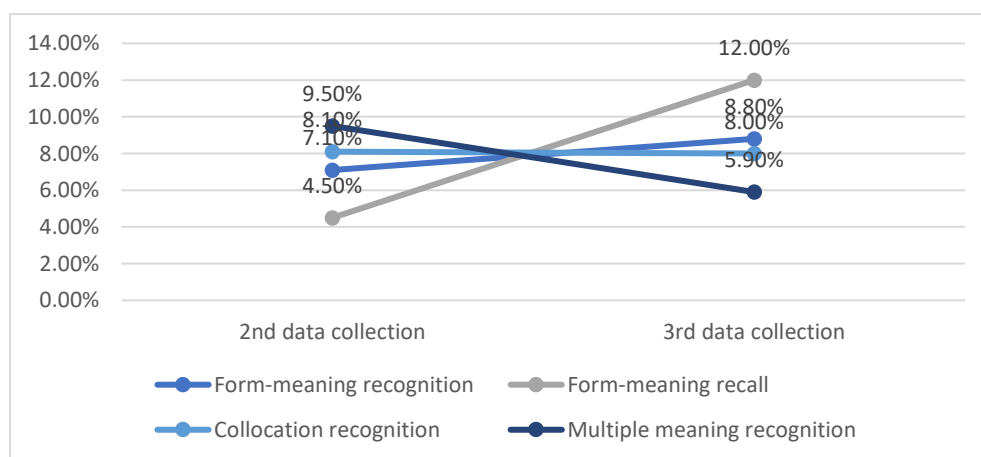


Figure 10.6: Development of total variance explained by EE predictor variables in a general population

Comparing collocation recognition knowledge to the two form-meaning knowledge components shows that this aspect of word knowledge appears to be more in line with form-meaning recognition, showing only a minimal change of 0.1% over time. In contrast, multiple meaning recognition knowledge displays a relatively clear decrease of variability explained by the EE factors.

10.2 Biographical background variables & word knowledge aspects considered in an EE setting

To further complement the search for predictor variables of the four word knowledge components in focus the influence of biographical background variables and word knowledge aspects considered in an EE setting have been examined. While no significant findings could be detected for the third data collection, the following results were revealed for the second data collection.

Analysing the relationship between the biographical variables gender, age of onset, age, number of languages spoken, experience abroad, extramural speaking opportunities and educational history and the four word knowledge aspects shows that only multiple meaning recognition knowledge appears to be in linear relationship with one biographical background variable (see Table 10.34).

	Form-meaning recognition	Form-meaning recall	Collocation recognition	Multiple meaning recognition
Gender	$r_s = -0.09, p = 0.34$	$r_s = -0.02, p = 0.88$	$r_s = -0.08, p = 0.42$	$r_s = 0.00, p = 0.998$
Age of onset	$r_s = -0.11, p = 0.28$	$r_s = -0.03, p = 0.76$	$r_s = 0.01, p = 0.92$	$r_s = -0.05, p = 0.63$
Age	$r_s = 0.07, p = 0.46$	$r_s = -0.08, p = 0.42$	$r_s = 0.15, p = 0.11$	$r_s = 0.03, p = 0.74$
Number of languages spoken	$r_s = 0.06, p = 0.52$	$r_s = 0.13, p = 0.17$	$r_s = 0.00, p = 0.989$	$r_s = 0.19, p = 0.05^*$
Extramural speaking opportunities	$r_s = 0.04, p = 0.68$	$r_s = -0.05, p = 0.59$	$r_s = 0.08, p = 0.41$	$r_s = 0.03, p = 0.77$
Experience abroad	$r_s = -0.03, p = 0.74$	$r_s = -0.11, p = 0.25$	$r_s = 0.04, p = 0.70$	$r_s = 0.06, p = 0.57$
Educational history	$r_s = -0.11, p = 0.25$	$r_s = -0.03, p = 0.80$	$r_s = -0.00, p = 0.98$	$r_s = 0.01, p = 0.95$

Table 10.34: Spearman correlation matrix biographical background variables & word knowledge components

Interpreting the correlation matrix reveals that when the number of languages spoken by participants increases multiple meaning recognition knowledge of academic verbs rises as

well. Hence, students who speak more languages appear to have higher scores on multiple meaning recognition knowledge.

To allow an interpretation of a causal relationship between these two variables a linear regression model has been computed, showing that multiple meaning recognition knowledge is significantly influenced by the number of languages spoken by the participants ($F(1,104)=4.943, p=0.028$). Overall, the number of languages spoken by participants appears to explain 4.5% of the variability in multiple meaning recognition knowledge in the given population and 3.6% in a generalized sample.

In a next step, students' consideration of different word knowledge aspects in their spare time were considered as potential predictor variables for the vocabulary knowledge of academic verbs. Nonetheless, a very similar picture as for biographical background variables can be observed when looking at Table 10.35 displaying all Spearman correlations.

	Form-meaning recognition	Form-meaning recall	Collocation recognition	Multiple meaning recognition
Spelling	$r_s=0.03, p=0.74$	$r_s=0.07, p=0.48$	$r_s=-0.01, p=0.92$	$r_s=0.06, p=0.54$
Oral repetition & pronunciation practice	$r_s=-0.14, p=0.17$	$r_s=-0.11, p=0.28$	$r_s=-0.08, p=0.42$	$r_s=0.00, p=0.96$
Phonetic transcription	$r_s=0.08, p=0.42$	$r_s=0.12, p=0.23$	$r_s=0.04, p=0.66$	$r_s=0.17, p=0.08$
Dividing the word into word parts	$r_s=-0.03, p=0.78$	$r_s=-0.11, p=0.29$	$r_s=-0.04, p=0.69$	$r_s=0.01, p=0.85$
Searching for words in other languages	$r_s=-0.10, p=0.32$	$r_s=-0.13, p=0.18$	$r_s=-0.11, p=0.27$	$r_s=-0.20, p=0.04^*$
Searching for members of a word family	$r_s=0.07, p=0.49$	$r_s=0.07, p=0.46$	$r_s=0.15, p=0.12$	$r_s=0.15, p=0.14$
Searching for translations	$r_s=-0.17, p=0.08$	$r_s=-0.18, p=0.06$	$r_s=-0.18, p=0.07$	$r_s=-0.24, p=0.02^*$
Searching for definitions	$r_s=0.13, p=0.18$	$r_s=-0.03, p=0.74$	$r_s=0.09, p=0.37$	$r_s=-0.02, p=0.81$
Checking in which different context the word is used	$r_s=-0.02, p=0.86$	$r_s=-0.10, p=0.31$	$r_s=0.00, p=0.98$	$r_s=-0.07, p=0.50$
Searching for different shades of meaning	$r_s=0.10, p=0.30$	$r_s=0.01, p=0.94$	$r_s=0.07, p=0.46$	$r_s=0.14, p=0.15$

Searching for synonyms	$r_s=0.04, p=0.70$	$r_s=-0.09, p=0.37$	$r_s=0.11, p=0.28$	$r_s=-0.12, p=0.23$
Searching for antonyms	$r_s=-0.02, p=0.87$	$r_s=-0.09, p=0.37$	$r_s=0.12, p=0.23$	$r_s=-0.01, p=0.93$
Using the word in a sample sentence	$r_s=0.05, p=0.64$	$r_s=-0.09, p=0.34$	$r_s=0.07, p=0.46$	$r_s=0.03, p=0.75$
Search for the word class	$r_s=0.03, p=0.76$	$r_s=0.07, p=0.48$	$r_s=0.07, 0.46$	$r_s=0.07, p=0.47$
Check correct grammatical use	$r_s=0.00, p=0.97$	$r_s=0.02, p=0.81$	$r_s=0.10, p=0.31$	$r_s=-0.07, p=0.45$
Search for frequent collocations	$r_s=-0.02, p=0.85$	$r_s=-0.08, p=0.41$	$r_s=0.00, p=0.93$	$r_s=0.03, p=0.71$
Check level of formality the word	$r_s=0.00, p=0.99$	$r_s=-0.05, p=0.60$	$r_s=0.05, p=0.61$	$r_s=-0.04, p=0.66$
Check currency of the word	$r_s=-0.00, p=0.98$	$r_s=-0.07, p=0.50$	$r_s=0.06, p=0.55$	$r_s=0.00, p=0.97$
Check meaning differences in different English varieties	$r_s=0.00, p=0.99$	$r_s=-0.15, p=0.13$	$r_s=0.05, p=0.59$	$r_s=0.02, p=0.81$
Consider level of abstraction of a word	$r_s=0.05, p=0.65$	$r_s=-0.08, p=0.44$	$r_s=0.01, p=0.91$	$r_s=0.11, p=0.26$
Check relation to specific domain	$r_s=-0.02, p=0.85$	$r_s=0.10, p=0.31$	$r_s=0.00, p=0.96$	$r_s=0.04, p=0.68$

* significant correlation for overall Multiple meaning recognition knowledge

Table 10.35: Spearman correlation matrix for consideration of word knowledge aspects for words encountered in EE activities

As with biographical background variables only multiple meaning recognition knowledge has a significant linear relationship with two variables. Firstly, the more students use the strategy to use other known languages to understand a new word in their spare time, the less multiple meaning recognition knowledge they appear to have. Secondly, a similar picture is found for students using translations extensively.

To investigate the causal relationship between the two spare time strategies and overall multiple meaning recognition knowledge a multiple linear regression model has been calculated. As visible in Table 10.36 searching for translations significantly ($p=0.047$) influences multiple meaning recognition knowledge, while searching for words in other languages loses its predictive power due to insignificant unique variance ($p=0.14$).

	B	95% CI	SE	t	p	β
<i>Constant</i>	21.832	[19.905, 23.759]	0.971	22.476	<0.001*	
Searching for translations	-0.482	[-0.958, -0.006]	0.240	-2.008	0.047*	-0.203
Searching for words in other languages	-0.215	[-0.498, 0.068]	0.143	-1.508	0.135	-0.153

Table 10.36: Model parameters multiple meaning recognition & Word knowledge aspects predictors

Taken together both predictors make a significant contribution to predicting multiple meaning recognition knowledge ($F(1,102)=4.846$, $p=0.01$) and are able to explain 0.9% of the variability in the given and 0.7% in a general sample.

Hence, while form-meaning recognition, form-meaning recall and collocation recognition knowledge appear to be solely influenced by EE activities, multiple meaning knowledge seems to be additionally affected by other variables. Combining all predictor variables of multiple meaning knowledge into one multiple linear regression model reveals that all of them make a significant contribution to explaining multiple meaning recognition knowledge ($F(5,98)=5.614$, $p<0.001$) as visible in Table 10.37.

Model	R	R²	Adjusted R²	Sig. F Change
Creative writing + Fiction-oriented reading	0.318	0.101	0.084	0.005
Creative writing + Fiction-oriented reading + Searching for translations + Searching for words in other languages	0.436	0.190	0.157	0.006
Creative writing + Fiction-oriented reading + Searching for translations + Searching for words in other languages + Number of languages	0.472	0.223	0.183	0.045

Table 10.37: Model summary linear regression multiple meaning recognition & Word knowledge aspects predictors

Together Creative writing, Fiction-oriented reading, searching for translations, searching for words in other languages and number of languages spoken are able to explain 22.3% of the

variability in the observed sample and at least 18.3% in a general population. However, having a look at Table 10.38 reveals that as in the individual models two variables lose their predictive power, namely Creative writing and searching for translations.

	B	95% CI	SE	t	p	β
<i>Constant</i>	20.703	[18.562, 22.843]	1.079	19.194	<0.001*	
Fiction-oriented reading	0.479	[0.085, 0.873]	0.199	2.410	0.018*	0.249
Creative writing	0.223	[-0.260, 0.705]	0.243	0.916	0.362	0.094
Searching for words in other languages	-0.308	[-0.586, -0.030]	0.140	-2.200	0.030*	-0.218
Searching for translations	-0.384	[-0.839, 0.071]	0.229	-1.673	0.097	-0.162
Number of languages	0.354	[0.009, 0.699]	0.174	2.034	0.045*	0.193

Table 10.38: : Model parameters multiple meaning recognition & all predictors

This means that in order to improve multiple meaning recognition knowledge students might want to engage in Fiction-oriented reading activities, such as reading English books or short stories. Moreover, students who speak more languages seem to have a benefit when it comes to multiple meaning recognition knowledge. However, surprisingly it should be rather avoided to search for words in other languages, since this appears to negatively impact multiple meaning recognition progress.

11 Discussion

Building on previous research of two main research fields this study addresses several questions, aiming to contribute to filling potential research gaps. In the area of vocabulary knowledge this study strives to enlarge research insights regarding academic word knowledge and provides longitudinal results on the development of four different word knowledge components in the context of Austrian Bachelor students majoring in English. Moreover, it was attempted to investigate the nature of the relationship between these four word knowledge components, adding to a new and still rather scarce number of studies looking into this topic (e.g. González-Fernández & Schmitt, 2019). Concerning research into Extramural English several studies have investigated the EE behaviour of students at tertiary level before (e.g. Busby, 2020; Kusyk, 2017; Liu, 2014). However, this study tries to contribute to these findings by adding information about the stability of EE behaviour over time.

Finally, linking the two research fields under investigation, this study provides insights into the relationship between Extramural English activities and academic vocabulary knowledge, expanding on foci of prior research concentrating mainly on vocabulary size. However, the study does not only allow insights into the relationship between specialized word knowledge and EE but further zooms in on the influence of EE on specific word knowledge aspects of academic verbs over time. Therefore, this study is additionally offering a unique longitudinal perspective providing first insights regarding the stability of links between EE and academic vocabulary knowledge.

The following chapter will briefly summarize, contextualize and interpret the main findings related to these research gaps through linking the results to prior findings and theories. Moreover, the theoretical and practical implications of this study as well as potential limitations will be outlined.

11.1 Developing vocabulary tests for academic verb knowledge

While the literature review has shown that several standardized vocabulary tests on general vocabulary size are available, the number of tests including (e.g. VLT) or solely examining academic word knowledge (e.g. AVT) is clearly more limited. Moreover, most of the available tests cover only one word knowledge aspect, while test batteries testing different word knowledge aspects of the same target words (e.g. González-Fernández & Schmitt, 2019; González-Fernández, 2021) are rare. However, when combining these two aspects no test batteries on academic words have been available at the time of methodology development of this study. Consequently, it was decided to develop a test battery covering academic verb knowledge of the four different word knowledge components investigated, which is target made for the participant group observed as promoted by Schmitt, Nation and Kremmel (2020, p. 111). All tests developed have been tested for their reliability and their suitability for the target group by looking at facility and discrimination values, as well as testing distractors for their efficiency in several piloting and data collection rounds.

At the same time one has to be aware of the fact that test development is a highly complex process, which has filled whole PhDs (e.g. Gyllstad, 2007). This implies that while the test instruments for this study have been developed faithfully as outlined in chapter 7, the tests might still have particular limitations that need consideration and more validation before being widely used or published.

Firstly, as mentioned by Schmitt, Nation and Kremmel (2020, p. 112) replicating existent test formats can mean a duplication of existent flaws. As already outlined in chapter 3.2 existent test formats have not been extensively tested for their reliability and validity either and already have some known flaws. Therefore, Schmitt, Nation and Kremmel (2020, p. 112) call for more creativity in test development to develop new test formats. Nonetheless, the test battery developed has been based on several existent test formats, such as the VLT (Nation,

1990; Schmitt, Schmitt & Clapham, 2001), the PVLТ (Laufer & Nation, 1999), the COLLEX (Gyllstad, 2009; Gyllstad, 2007) or the WAT (Read, 1993, 1998) due to their wide-spread use and out of practicability reasons. However, the test formats have been adapted and validated for the target group as outlined in chapter 7, which has counteracted some of their prior known limitations. For example the Form-meaning Recall Test of Academic Verbs has adopted the basic item format of the PVLТ (Laufer & Nation, 1999), diverging from the traditional format concerning the context reduced sentence context in which the target words are presented, which should prevent test takers from guessing answers from context. Furthermore, the intervention of reading skills and vocabulary knowledge of the sentence context was reduced to a minimum by creating a lexical profile of the instructions and the sentence context, which ensures a coverage of 95% within the most frequent 3000 words of the English language. Lastly, the original formats provides test takers with as many letters of the target word searched for as necessary to prevent ambiguity. This procedure has been criticised heavily in the past (Read, 2010), since test items are of varying difficulty due to the different number of letters given. This was prevented by adding a translation element to the test. While only two letters are given for each target word ensuring similar difficulty of test items, a German translation of the word is given beneath the sentence context to ensure unambiguity of answers. This additionally weakens the necessity of guessing answers from context supporting a clearer construct measured. Similar precautions and adaptations have been made for the other three test formats developed to counteract existent flaws. However, future studies definitely should try to validate the existent tests further but should also go beyond these steps and develop other target made test formats suited to new insights of vocabulary research.

Secondly, one common limitation in all vocabulary tests is the problem of inflation of scores through guessing (Gyllstad, Vilkaite & Schmitt, 2015; Schmitt, Nation & Kremmel, 2020, p. 114). While this is also an issue for this test battery a first attempt to counteract guessing has been made in the Collocation Recognition Test of Academic Verbs by offering the participants the option to show-case when they had to guess. This is similar to the principle of providing participants with an “I don’t know” option, which has shown to improve guessing behaviour and exactness of the test score (e.g. Peters, Velghe & Van Rompaey, 2019; Stoeckel, Bennett & McLean’s, 2016). As mentioned above the Form-meaning Recall Test of Academic Verbs tried to counteract guessing from context by reducing the information in the sentence context given and adding a translation aspect to decrease the perceived necessity to guess. However, there is still room for further improvement, since the other test formats did work with distractors but included no other means to counteract guessing or at least make guessing behaviour visible apart from administrative steps taken, such as informing participants that the test should reflect their true knowledge of the words and that test results have no consequences for them.

Moreover, Schmitt, Nation and Kremmel (2020, p. 117) outline that for newly developed tests it has to be made clear what knowing a word on the test means for language ability. While it can be assumed that the tests of recognition knowledge are linked to the skill of reading and the Form-meaning Recall Test of Academic Verbs to writing, this link and its strength needs to be measured in a further validation step. Therefore, more reflection will be necessary concerning the usability of the test formats for other language abilities.

This is closely linked to the use of the notion of word frequency for test development. As Kremmel (2016) and Kremmel and Schmitt (2016) outline most tests use word frequency to extrapolate the total vocabulary size of a person based on their knowledge of individual words. Although this is common practice it has been criticised, since word frequency is not the only factor influencing vocabulary acquisition (e.g. Kremmel, 2016; Kremmel & Schmitt, 2016). Consequently, while this test battery uses frequency estimates to sample test items, no claims beyond the knowledge of these items have been made.

Furthermore, while all tests included in the test battery have been tested for reliability by using Cronbach alpha as a measure for internal reliability, which has been widely in use, Schmitt, Nation and Kremmel (2020, p. 115) state that future test validation should rather rely on test-retest procedures (Schmitt, 2010) or the concept of summability⁸⁹. As Cronbach alpha as a measure of internal reliability assumes that all test items in a test measure the same thing, such as the plural s in a grammar test, this is not 100% suitable for vocabulary tests, where each item measures knowledge of a different word (Schmitt, Nation & Kremmel, 2020). However, also here the whole vocabulary test should picture one construct, being based on the word knowledge aspect tested or the word type. Nonetheless, Cronbach alpha has also been shown to be sensitive to item number, with a tendency to be exaggerated for tests with a high item number. Therefore, test-retest procedures and measures of summability will have to be calculated for the test battery in a next step to further improve their reliability and validity.

Finally, Kremmel (2020) and Nation and Webb (2011) highlight problematic issues regarding study designs measuring vocabulary development. Firstly, they state that all tests administered need to be identical regarding items, test format and administration. While test items stayed the same over the time of investigation, administration, while being systematic and similar overall, varied slightly due to more experience gained over time as described in chapter 7. Moreover, while the test format stayed the same for the Form-meaning Recognition and Form-meaning Recall Test, as well as for the Multiple Meaning Recognition Test, it was varied for the Collocation Recognition Test, using once a format related to the WAT (Read, 1993, 1998) and once a format similar to the COLLEX (Gyllstad, 2007). However, correlational analysis showed that participants performed similarly on the

⁸⁹ The concept of summability is described as the “proportion of total test variation that is explained by the sum score” of the test by Goeman and de Jong (2018, p. 54).

two tests. Whereas the test format was changed for the Collocation Recognition Test to counteract test effects as mentioned by Nation and Webb (2011), the item order was changed for the other tests, to distract learners from the fact that the same 72 words were tested. Another option which might be adapted in future versions of the tests, would be to include items in the tests which are not part of the investigation to distract learners from the target items (Nation & Webb, 2011). Yet, this option would require more time to administer the tests again, which was not available within this study design.

11.2 Academic verb knowledge at tertiary level in Austria

When examining the scores on the four vocabulary tests taken the overall high performance of participants already at the start of the study, with a median of 60 points out of 72 points for the Form-meaning Recognition Test and a mean of 35.34 points out of 72 points for the Form-meaning Recall Test, becomes visible.

Looking at students' correct answer rate of 83.3% on the first Form-meaning Recognition Test reveals that students appear to have been very proficient regarding this word knowledge component from the start. This high result could indicate that the test developed has been too easy for the participants from the beginning, despite different piloting results. However, when comparing students' performance on the Form-meaning Recognition Test to prior studies reveals comparable results using different test instruments. Considering performance on form-meaning recognition tests on general vocabulary size, like the VLT, participants' results are comparable to the findings reported by Ma and Lin (2015, p. 7) with 88% and Staehr (2009, p. 590) with 82.44% correct answers, being the two studies reviewed with the best results. Zooming in further on research studies on academic vocabulary specifically shows a similar picture. Both Ozturk (2015) and Tschirner (2004), investigating test performance at tertiary level, describe high results on academic vocabulary tests with 88% (Ozturk, 2015, p. 9) and over 76% of the participants reaching over 80% correct answers (Tschirner, 2004, p. 31).

Furthermore, the particularly high results on the Form-meaning Recognition Test might be attributed to the fact that this word knowledge component is amongst the first aspects of word knowledge acquired (González-Fernández & Schmitt, 2019, p. 13) and usually reveals higher scores than productive counterparts (e.g. Cheng & Matthews, 2018; Daller, Milton & Treffers-Daller, 2007; Lee, 2019a; Miralpeix, 2020). This is also reflected in this data set with students only reaching an average of 49.08% correct answers on the Form-meaning Recall Test at the first data collection. This falls into the range of 40%-70% in which most findings of the studies reviewed fall (e.g. Fan, 2000; Nizonkiza, 2016; Uchiyara & Saito, 2019).

Moreover, taking into account participants' profile, being English major students who have passed the school leaving examination with a B2 language level relatively recently implies that high results might be partly justified by a relatively high overall language proficiency.

This is further supported by the fact that all students participating had to take a standardized language test including vocabulary knowledge above B2 level to be permitted to study English (VELT, 2022). This is further supported by Lee (2019a, p. 773), who reports a significantly better receptive and productive vocabulary size of English major students in comparison to non-English major students. Hence, an overall high level of vocabulary knowledge as visible in the results of the Form-meaning Recognition and Recall tests might be expected for this participant group. This is also reflected in the results of the Collocation Recognition Test and Multiple Meaning Recognition Test.

While most studies on collocation recognition knowledge reviewed have reported a correct answer rate between 40%-60%, the participants of this study reached an average of 65.38% (*Mdn*=17 out of 26 points). However, similar results have been found by Wang (2014) with 67.5% correct answers on the WAT and Qian (2002) with 70.66% correct answers on the DVK. Furthermore, other studies revealed even higher results ranging from 72.06% (Talakoob & Koosha, 2017, p. 14) to 96.13% (Kamarudin, Abdullah & Abdul Aziz, 2020, p. 4).

Regarding multiple meaning recognition studies mixed findings have been reported, which might be loosely linked to the educational level investigated. While studies on multiple meaning recognition at secondary school level revealed relatively high results, studies at tertiary level revealed mixed findings ranging from 31.2% (Alnamer, 2017) to 67.2% (Zhang & Wen, 2019) correct answers. Consequently, the participant group of this study appears to outperform these studies by reaching 79.17% on the test (*Mdn*=19 out of 24 points). This becomes even more surprising regarding the fact that word and sense-based frequency appear to significantly impact students' performance on multiple meaning recognition tests (e.g. Karlsson, 2013; Maby, 2016; Sonbul, El-Dakhs & Al-Otaibi, 2020), with less frequent words and senses being harder to answer for students. While an impact of word or sense-based frequency within the Multiple Meaning Recognition Test used has not been examined within the scope of this study, the overall finding does not correspond to the assumption that test results on a multiple meaning recognition test of academic verbs might be more difficult than a test on general high-frequency words. However, it has to be considered that the Multiple Meaning Recognition Test used in this study comprises only twelve academic verbs, which might lead to comparability difficulties with other research instruments, such as acceptability judgement tests or the Multiple meaning test (Qian, 2002).

11.3 Development of academic verb knowledge

Moving on to the question whether participants improved their vocabulary knowledge, it can be stated that a significant development was observed for all four word knowledge components in one academic year. This corresponds to findings of other studies like Yamamoto (2014), Ozturk (2016) and Gyllstad (2004). However, when considering the effect size of the developments it is visible that form-meaning recall knowledge appears to have

developed considerably more than form-meaning recognition knowledge with an η^2 of 0.50. In addition, while the effect sizes⁹⁰ of r and η^2 are not directly comparable it can still be assumed that form-meaning recall knowledge is followed by multiple meaning recognition knowledge, collocation recognition knowledge and only finally form-meaning recognition knowledge concerning the growth rate (visible in Figure 11.4). This corresponds to findings by Webb (2005, 2007) and Schmitt (1998), who report a parallel development of their word knowledge components observed but at slightly different growth rates. In a pioneering study González-Fernández and Schmitt (2019) take the investigation of four word knowledge components even one step further. By using Guttman scaling (Wichmann, 2015) they reveal that when they order the components according to their level of mastery this can be assumed to be a suitable implicational scale (González-Fernández & Schmitt, 2019, p. 13) looking as follows:

Form-meaning recognition > Collocation recognition > Multiple meaning recognition > Derivative recognition > Collocation recall > Form-meaning recall > Derivative recall > Multiple meaning recall

Comparing their findings roughly to this study's results, which have not been checked with Guttman scaling (Wichmann, 2015), we still find a fairly similar result⁹¹:

Form-meaning recognition > Multiple meaning recognition > Collocation recognition > Form-meaning recall

González-Fernández and Schmitt (2019, p. 13) further interpret these findings as evidence that for example form-meaning recognition knowledge is easier than collocation recognition knowledge and that it might be assumed that someone who knows the collocations of a word receptively is capable of knowing the form-meaning recognition of a word. However, one has to be careful not to interpret this implicational scale as a universal acquisition order, with form-meaning recognition knowledge being acquired first, followed by collocation recognition knowledge and all other word knowledge components measured. While this interpretation is intriguing, it still has to be considered that this implicational scale is originally based on test scores giving information about the potential ease of each task at one point in time⁹² and that several factors might influence acquisition (González-Fernández

⁹⁰ r = The Pearson correlation coefficient ranges between -1.0 and +1.0 and can be interpreted as effect size. It provides information about the strength of a connection between two variables (McLeod, 2019).

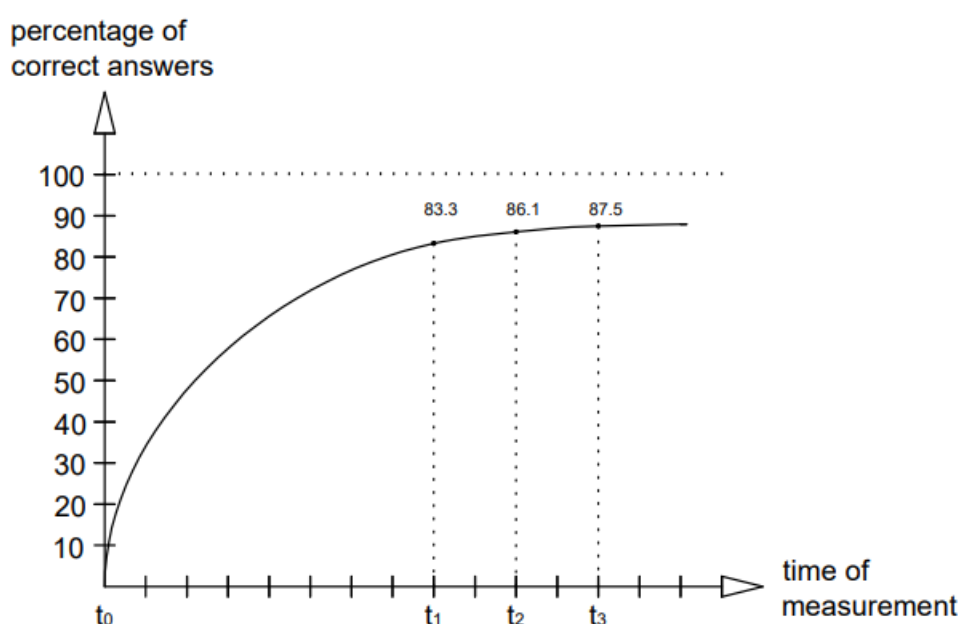
η^2 = "Eta-squared is commonly used in ANOVA and t test designs as an index of the proportion of variance attributed to one or more effects. The statistic is useful in describing how variables are behaving within the researcher's sample. In addition, because eta-squared is a measure of effect size, researchers are able to compare effects of grouping variables or treatment conditions across related studies." (Fay & Boyd, 2010)

⁹¹ In a further step Guttman scaling could be used to make results more comparable.

⁹² The difference between González-Fernández and Schmitts' (2019) study and this study is that their implicational scale is based on data collected at only one point in time. Consequently, their implicational scale is based on participants' level of mastery of the word knowledge components. In contrast, statements about

& Schmitt, 2019, p. 18). Therefore, more longitudinal validation studies need to be carried out to confirm and place the interpretation of this implicational scale as potential acquisition order.

Trying to add some more insights regarding the longitudinal development of different word knowledge components it appears to be very interesting that the most difficult word knowledge component measured, namely form-meaning recall knowledge, developed the most in the time of investigation (see also González-Fernández & Schmitt, 2019, p. 13). In comparison the easiest word knowledge component observed, namely form-meaning recognition developed considerably less with an η^2 of 0.21. Therefore, an interpretation of the development of the word knowledge components as consistently linear appears to be unlikely, since prior research shows that recognition knowledge continuously outweighs recall knowledge (e.g. González-Fernández & Schmitt, 2019; González-Fernández, 2021; Nizonkiza, 2015). This appears to be further supported by research indicating that acquiring recall knowledge seems to be more difficult than recognition knowledge (e.g. González-Fernández & Schmitt, 2019, p. 18, Pellicer-Sánchez & Schmitt 2010; Van Zeeland & Schmitt 2013). Instead, a developmental curve as depicted in Figure 11.1 for form-meaning recognition knowledge appears to be more likely.



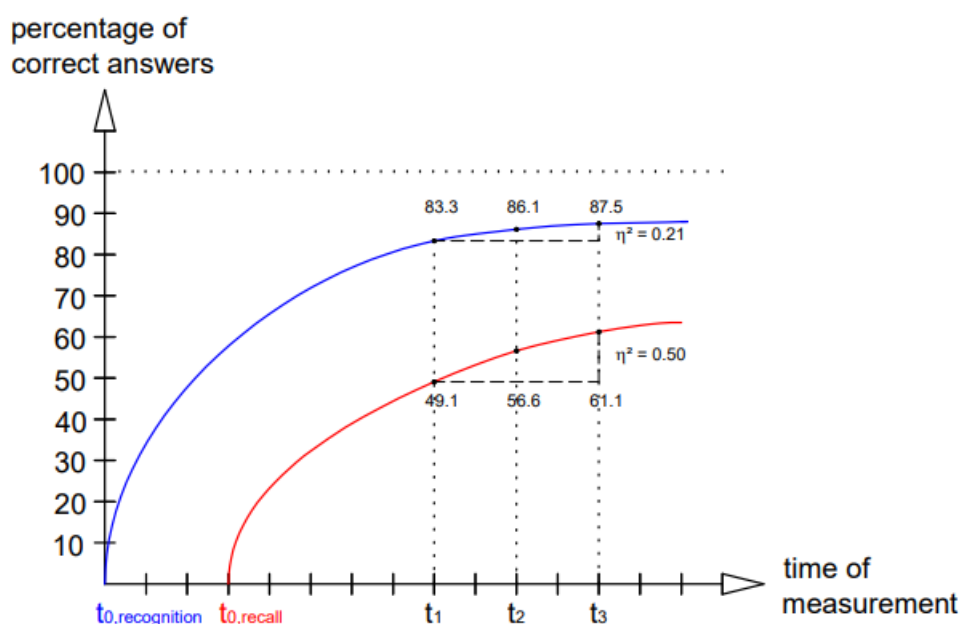
Note: While the development between t₁ and t₃ is based on the results of this study, the development of the curve between t₀ and t₁ is an estimate.

Figure 11.1: Developmental curve form-meaning recognition knowledge

In the line of Bordag, Gor and Opitzs' (2022) ontogenesis model word knowledge gain is described as a curve, which flattens towards the end. Therefore, the rather small growth

an implicational scale in this study are based on the development of academic verb knowledge over three points in time and not only on mastery of word knowledge components at one point in time.

rate of form-meaning recognition knowledge in the study, encompassing three times of measurement (t_1 - t_3), would be explained by the fact that students were already quite proficient from the start and have reached a form of plateau with slower growth rates. This interpretation further presumes that early knowledge gain is easier than closing the gap between already proficient knowledge of a word knowledge aspect to full knowledge. This is also supported by the larger growth rate of form-meaning recall knowledge during the time of investigation (see Figure 11.2) or the other word knowledge components (see Figure 11.4).



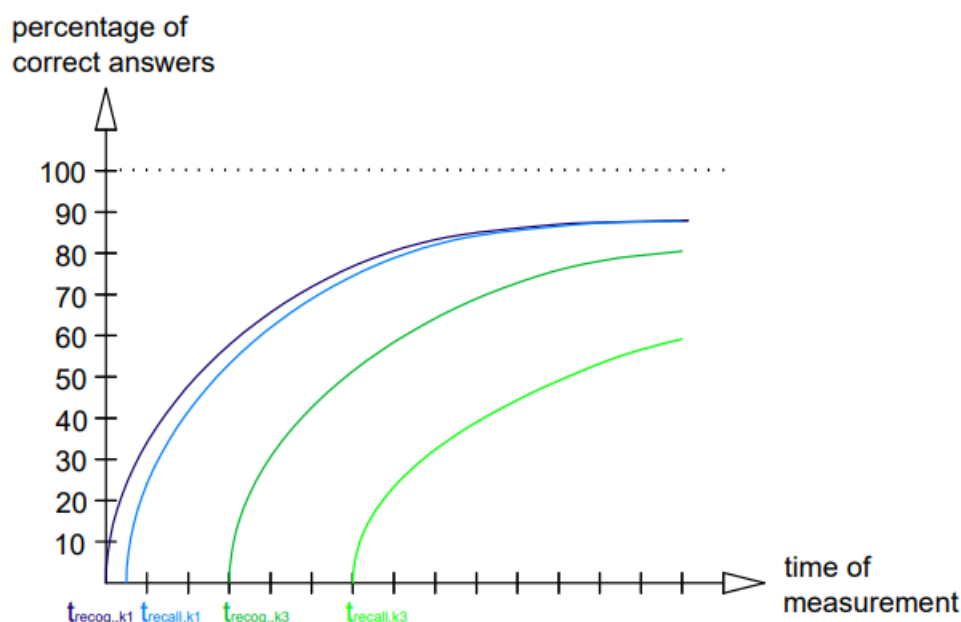
Note: While the development between t_1 and t_3 is based on the results of this study, the development of the curve between t_0 , recognition or t_0 , recall and t_1 is an estimate.

Figure 11.2: Developmental curves and growth rates form-meaning recognition and recall

As visible in Figure 11.2 two more aspects might come into play when trying to explain the varying growth rates. Firstly, as proposed by González-Fernández and Schmitt (2019, p. 19) word knowledge aspects, such as form-meaning recognition and recall depicted in Figure 11.2, might be acquired at different points in time (t_0 , recognition, t_0 , recall), with recall knowledge tending to start later than recognition knowledge due to its increased difficulty. Therefore, form-meaning recall knowledge might be at a different developmental stage than form-meaning recognition knowledge at the points of measurement (t_1 - t_3) explaining the larger growth. This would additionally correspond to Bordag, Gor and Opitzs' (2022) conceptualization of the linguistic domains in their ontogenesis model⁹³, where they state that different components might emerge at different times. However, it appears to be relatively unlikely that the difference in knowledge and development might be solely explained by the order of acquisition. As visible in Figure 11.3 it appears to be reasonable

⁹³ See chapter 2.4.2.5 for further information on the ontogenesis model by Bordag, Gor & Opitz (2022).

that the order of acquisition and the gap between the onset of different word knowledge aspects might differ according to frequency levels.



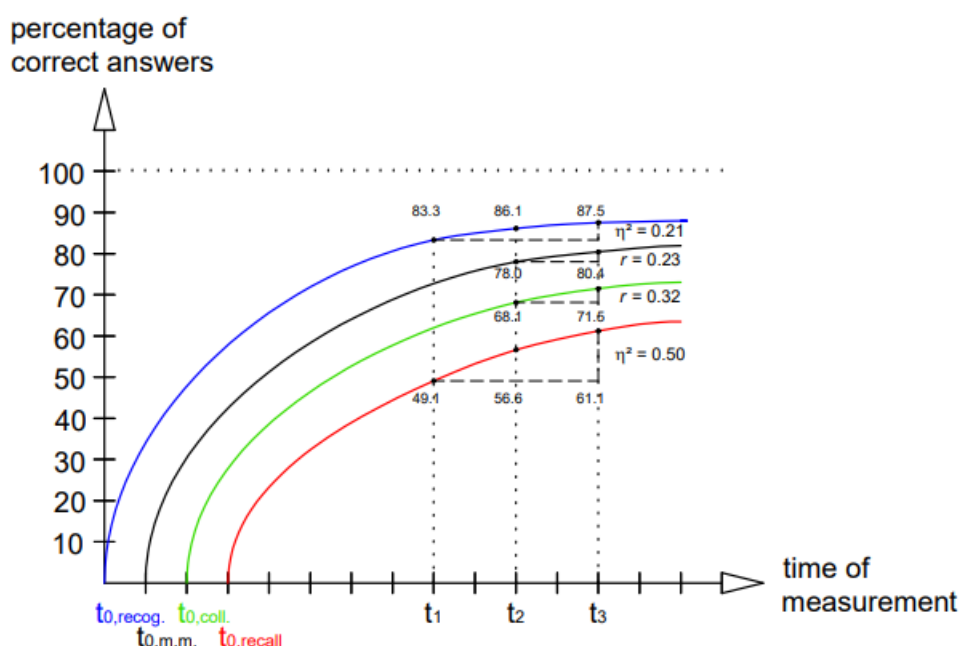
Note: The curves depicted in this graph are an estimate.

Figure 11.3: Developmental curves dependent on frequency of use

As visualized in Figure 11.3 it is hypothesized that words of a higher frequency band visualized in blue ($k1$ =first 1000), such as the word *year*, are not only learned receptively relatively early but might also be transferred into productive knowledge soon after the first encounter. In contrast, less frequent words visualized in green, such as *retirement* ($k3$ =3000 most frequent words) are not only acquired at a later stage but also take longer to move from form-meaning recognition to recall. This is supported by several studies (e.g. Brysbaert et al., 2016; Coxhead & Boutorwick, 2018) showing that word frequency has a significant, however not unlimited (see for example Brysbaert, Mandera & Keuleers, 2018), impact on word knowledge. For instance, it has to be considered that high-frequency lists are generated from corpora for an average language user. However, individualised needs influenced by factors such as profession, Extramural English use or a semester abroad might alter which words are to be considered as high-frequency words for an individual person. For example, while the word *surface friction* can be considered to be a (technical) low frequency word for an average language user, a structural engineer will need this word frequently in his/her everyday life. The same goes for the word *revenge*, which can be allocated to the 5000 most frequent words in the English language but might be much more frequently used by players of the online game World of Warcraft. Moreover, research has shown that the effect of frequency is larger for lower proficiency students than high proficiency students (Brysbaert, Mandera & Keuleers, 2018). Consequently, it will be

important in the future to undertake detailed case studies looking into the English needs and the related vocabulary development.

Additional to differences in the acquisition of different word knowledge aspects according to word frequency, it appears to be more likely that both order of acquisition and slightly different developmental curves dependent on the difficulty level of word knowledge aspects as depicted in Figure 11.4 below for all word knowledge components measured in this study come into play.



Note: While the development between t_1 and t_3 is based on the results of this study, the development of the curves between t_0 , recogn., t_0 , recall, t_0 , m.m. or t_0 , coll. and t_1 are an estimate.

Figure 11.4: Developmental curves for all word knowledge components measured

This is again supported by the ontogenesis model, which proposes different developmental curves for different components (Bordag, Gor & Opitz, 2022, p. 187)⁹⁴. Based on these hypothesis future studies should look into the long term development of various word knowledge components trying to reveal empirical insights regarding developmental curves and the role of acquisition order, word frequency and difficulty of word knowledge components, combining various methodologies.

11.4 The nature of vocabulary knowledge

Another area of vocabulary research this study is contributing to is to investigate the nature of vocabulary knowledge. As Schmitt (2019, p. 262) states future research studies should strive to “[develop] a practical model of vocabulary acquisition”. This has been attempted by adopting a longitudinal research design and utilizing SEM additional to correlational

⁹⁴ This study provides first empirical evidence for the ontogenesis model of (Bordag, Gor & Opitz, 2022).

analysis to investigate relationships between the components. Only few prior studies have attempted to measure several word knowledge aspects at a time (e.g. Chui, 2006; Schmitt, 1998; Webb, 2005; Webb 2007) and most of them have compared components via correlational analysis (e.g. Chui, 2006; Webb, 2005; Webb, 2007).

Looking at these binary relationships between the word knowledge components researched shows that regarding the relationship between form-meaning recognition and recall knowledge the findings of this study can be ranked with prior findings. These range from a large ($r=0.84$) to a medium ($r=0.41$) correlation between the receptive and productive knowledge of the form-meaning link (Cheng & Matthews, 2018; Miralpeix & Muñoz, 2018). While the correlation is smaller than Cheng and Matthews' (2018) finding, with $r=0.73$, $p<0.001$ at the most, it can still be interpreted as a large correlation.

A similar picture can be observed for the relationship between form-meaning recognition knowledge and collocation recognition knowledge. This study aligns with other studies reporting a medium correlation (Lee, 2020; Nguyen & Webb, 2017; Qian, 2002; Wang, 2014). In contrast other studies like Vafaei and Suzuki (2020) or Staehr (2009) have found large correlation coefficients for these two word knowledge components. On the basis of such different correlational strengths reported by different studies Enayat and Amirian (2020) propose that correlational strength might be indirectly proportionally related to the overall language proficiency of participants. However, this is called into question by the result of this study, since English major students, who can be considered to be advanced learners of English, should therefore display no significant relationship between the two word knowledge components, which is not the case for this study.

Moreover, also the connection between form-meaning recognition and multiple meaning recognition knowledge fits into this pattern with the correlation found in this study being in between prior results. While Lawrence et al. (2019) and Binder et al. (2017) report relatively high correlations of 0.63-0.67, Logan and Kieffer (2021) and Li and Kirby (2015) reveal medium sized to small correlations between 0.47-0.35. Therefore, the correlation coefficient of 0.54 at the most is located in between these extremes.

A different picture emerges for form-meaning recall knowledge and its connection to collocation recognition knowledge. Similar to prior studies, such as Zang and Lu (2015) and Zhong (2018) this study found a medium sized correlation between the two word knowledge components researched.

Building on these comparable findings, another interesting and novel aspect observed in this study is the stability of correlation strength over time. As visible in Table 11.1 correlation coefficients stay relatively stable over time, with the largest difference being 0.05.

	T1	T2	T3
Form-meaning recognition & Form-meaning recall	0.73**	0.68**	0.73**
Form-meaning recognition & Multiple meaning recognition	-	0.52**	0.54**
Form-meaning recognition & Collocation recognition	-	0.65**	0.62**
Form-meaning recall & Multiple meaning recognition	-	0.47**	0.47**
Form-meaning recall & Collocation recognition	-	0.50**	0.47**
Multiple meaning recognition & Collocation recognition	-	0.46**	0.46**

Table 11.1: Correlation coefficients for all word knowledge components

This is interesting not only on its own, but also when considering the mapping dimension in Bordag, Gor and Opitzs' (2022, p. 181) ontogenesis model (see Figure 11.5 again).

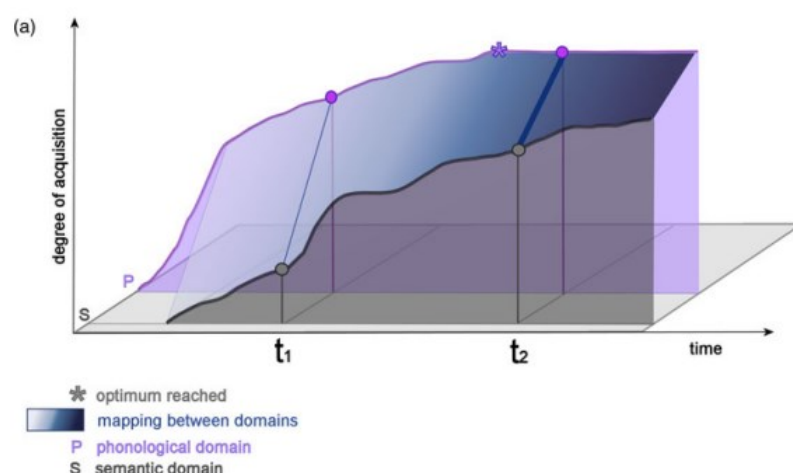


Figure 11.5: Illustration of the dimension of mappings taken from Bordag, Gor & Opitz (2022, p. 188)

As expanded in chapter 2.4.2.5, this dimension is concerned with connections between components, assuming that if components are connected more strongly, retrieval of the lexical representation should become easier. Consequently, when seeing correlations as a means to visualize connection strength between components, increased recall knowledge should go hand in hand with larger correlation values between word knowledge dimensions. Nonetheless, while there has been a significant positive development of form-meaning recall knowledge, correlation values remained fairly stable over the time of investigation. Yet, it might be possible that the combination of the limited time span observed in this

study and the fact that participants observed already display advanced knowledge for all word knowledge components, has left little room for observing these changes in the data.

Furthermore, this research study was not specifically designed to provide evidence or identify ambiguity in the ontogenesis model, which results in the fact that investigative measures used for the study might not be the most suitable to fulfil this purpose. For example, test formats on form-meaning recognition knowledge as found in prominent tests like the Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001) or the test used for this research study require participants to recognize a word form given and link it to a basic meaning provided through a definition. While such tests clearly measure some link between orthographic and semantic domain used in the ontogenesis model, the question remains which developmental stage of the ontogenetic curves of the two domains this test reflects and which degree of strength of the link between the domains is demonstrated through taking the test. Hence, current tests need to be thoroughly reflected regarding their role in the ontogenesis model and new, very sensitive tests need to be developed, allowing an examination of knowledge development.

Taking one step further in exploring the relationship between the different word knowledge components observed, this study offers a model (see Figure 8.11) how these individual aspects are related to the latent variable of overall vocabulary knowledge by using SEM modelling. Similar to the only other directly comparable SEM study (González-Fernández & Schmitt, 2019) the analysis has revealed that recognition and recall of each word knowledge type (e.g. form-meaning link) should rather be considered as individual components adding to overall vocabulary knowledge instead of two types of the same overarching knowledge type. This is further supported by González-Fernández (2021) who reports that recall and recognition knowledge of individual word knowledge components were statistically significantly different from each other. However, whereas González-Fernández and Schmitt (2019, p. 5) and González-Fernández (2021) focused on knowledge of twenty words coming from different frequency bands, this study has a narrow focus on academic verb knowledge. Therefore, it seems that the role of recognition and recall knowledge as separate constructs is not only generalizable across different populations but also different types of words and frequency levels. Moreover, in this line another congruent observation has been made, namely that form-meaning recognition knowledge contributed more to the overall construct than form-meaning recall knowledge.

Taking these findings into consideration when reviewing Nation's (2013) taxonomy of word knowledge reveals that all four word knowledge types measured are distinctive of each other. However, it cannot be stated with certainty whether the overarching category of form-meaning link knowledge split into receptive and productive knowledge should be omitted, since form-meaning recognition and recall knowledge were both explained to a satisfactory extent by this overarching category. This means that the overarching category

of form-meaning link knowledge was not omitted due to an unsatisfactory fit but due to its overlap with the overall concept of vocabulary knowledge. Consequently, future research studies should attempt to measure the receptive and productive side of more word knowledge components to give more satisfactory answers to these remaining questions.

While the overall SEM models arrived at several similarities with other research findings, finer differences can be observed when looking at the SEM models in more detail. González-Fernández and Schmitt (2019) found very high factor loadings for all their word knowledge aspects measured, whereas this study revealed rather varied factor loadings ranging between 0.89 and 0.55. Moreover, González-Fernández and Schmitt (2019) interpret their high factor loadings as an indication that all components researched can be explained by one unique factor, namely vocabulary knowledge. However, this study reached a different conclusion: Whereas a large proportion of form-meaning link knowledge, both receptive and productive is explained by the latent construct of vocabulary knowledge, multiple meaning and collocation recognition knowledge are explained less by this construct. This has already been foreshadowed by interpreting the correlation coefficients before as well. Hence, falling back on the classical distinction of vocabulary size (e.g. Anderson & Freebody, 1981, p. 93; Staehr, 2009, p. 578) and vocabulary depth (e.g. Qian, 2002, p. 515) it seems as if word knowledge aspects traditionally associated with vocabulary depth are explained by additional factors.

This is confirmed by the fact that multiple meaning recognition knowledge is not only predicted by EE activities, as all other word knowledge components researched, but is additionally explained by the number of languages spoken by participants and particular word knowledge aspects considered when meeting a new word. As the multiple linear regression model has shown the number of languages spoken by participants positively predicts multiple meaning recognition knowledge, which might be explained by an increased likelihood of meeting different meanings of a word in different languages. For instance, the word *bank* in English and the word *Bank* in German, share one identical referent, namely the institution where money can be deposited. Additionally, in English *bank* can be used to refer to a river bank or shore. A related meaning is also found in German with *Bank* being combined with a material like *Sand* to form *Sandbank* also related to shores. However, another meaning of *Bank* can be found in German, referring to a bench or seating accommodation. Consequently, having more languages to compare at one's disposal might aid multiple meaning recognition acquisition in the English language.

Interestingly, the influence of strategies applied when encountering a new word, considering translations or searching for words in another language implies the exact opposite. Based on the multiple linear regression model searching for translations or words in another language, which are clearly related, negatively predicts multiple meaning recognition knowledge. This means that students, who search for a translation or similar

words in another language when encountering a new word in their spare time, get a worse score on multiple meaning recognition knowledge. However, while this at first sight appears to contradict the positive influence of knowing multiple languages, it might be the case that those languages known are not always used most effectively when making one to one translations. For instance, when returning to the example of *bank* and *Bank* in English and German, German speaking participants might easily fall for the assumption that *Bank* in German has similar multiple meanings as *bank* in English. Therefore, using translations and words in another language without doing research on differences regarding meaning referents might lead to unsatisfactory knowledge of multiple meaning recognition knowledge.

While these findings give a first insight into potential other factors additional to vocabulary knowledge explaining multiple meaning recognition knowledge, more research studies will have to be conducted to investigate more closely how the variance in word knowledge aspects related to vocabulary depth might be further explained.

Further incorporating the longitudinal information gathered, adding information about the stability and durability of the SEM model, reveals that both the model and the similarities and differences observed remained stable over the course of one semester. While it might not seem surprising that the nature of vocabulary knowledge did not change drastically enough to suggest another model after one semester, the two models are also very similar regarding details, such as individual factor loadings. Both the size of the factor loadings as well as their comparable contribution to the construct of vocabulary knowledge remain fairly stable. Only collocation recognition knowledge appears to be even less explained by the latent construct of overall vocabulary knowledge over time, with the factor loading sinking from 0.71 to 0.64. A potential explanation for this decrease in explained variance over time might be that participants have developed more content knowledge during the time of investigation, which in addition to vocabulary knowledge helps them to identify correct and suitable collocations in an academic context but is not covered in the vocabulary test. This can be aligned to arguments of Coyle & Meyer (2021) pointing at the integrative nature of language and content in literacy development.

While this study is able to provide some interesting insights into the conceptualization of vocabulary knowledge and has been designed carefully, trying to avoid some limitations mentioned by González-Fernández and Schmitt (2019, p. 22), this study also has a similar limitation when it comes to measuring only a limited number of word knowledge components. As visible in Nation's (2013) taxonomy there are considerably more aspects one can know about words. However, measuring Nation's (2013) taxonomy as a whole would run into practical constraints considering the effort and time necessary to compile such a study. This still leaves a problem to be solved by future studies attempting to provide a more holistic picture of vocabulary knowledge.

11.5 Extramural English behaviour of Austrian English major students

Moving on to the second construct investigated in this study, namely Extramural English exposure this study tried to combine the separate items asking for individual activities in a meaningful manner instead of looking at the items individually or solely based on theoretical reasoning as prior studies have partially done (e.g. Chan, Spratt & Humphreys, 2002; Inozu, Sahinkarakas & Yumru, 2010; Littlewood & Liu, 1996; Sockett & Toffoli, 2012). For this purpose it was planned to run an exploratory factor analysis or a Principal components analysis on the whole questionnaire and use the computational variables as a means to further interpret and group the data. Unfortunately, though, no full exploratory factor analysis or Principal components analysis was possible for the whole questionnaire due to an unsatisfactory variable to case proportion of the data. Consequently, exploratory factor analyses were run on the theoretically conceived factors of the questionnaire individually to reveal whether each of them can be perceived as one component or whether those have to be split into several sub-categories. For this purpose the variables based on 5-point Likert scales were interpreted as interval data due to them having a clear mid-point, as frequently done in other linguistic fields. While not equalling the statistical rigour of a Principle components analysis, these individual factor analyses still revealed more of the underlying complexity of EE and the functionality of the questionnaire developed. Based on this analysis it can be proposed that skill-based categorization, such as reading activities in English or writing activities in English, might be too broad to capture the diversity of EE behaviour underlying these concepts. Instead, a distinction into Fiction-oriented reading, Information-oriented reading and Communication-oriented reading, as well as Creative writing and Written communication was used. Therefore, EE reading and writing activities seem to be influenced by the genre consumed, with Fiction-oriented reading including classical reading for pleasure, Information-oriented reading being focused on extraction of information from text and Communication-oriented reading being focused on the receptive part of communication. For the skill of writing Written communication comprises the productive communication with another person in a written manner and Creative writing the production of for pleasure in the spare time.

In a similar but considerably larger study (N=906) Sundqvist and Uztosun (under review) have investigated the underlying constructs of EE exposure using several different steps as visible in Figure 11.6 (see Sundqvist & Uztosun, under review, p. 2).

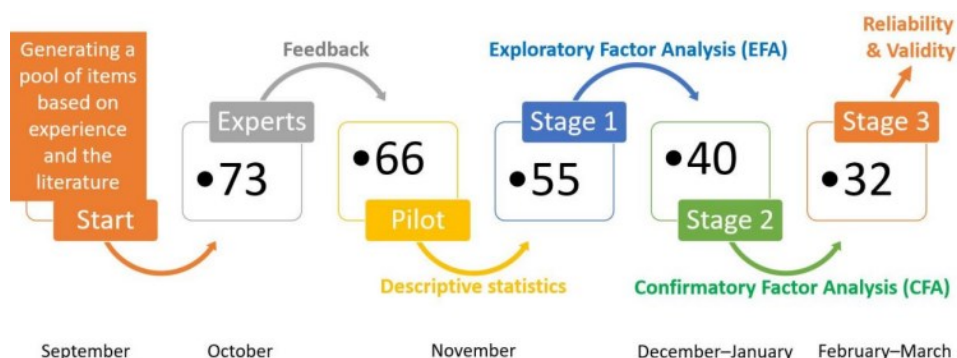


Figure 11.6: Development and Validation Procedures of the EE Scale Sundqvist & Uztosun (under review)

After generating a pool of 73 items they reduced them to 55 items considering expert feedback and descriptive statistics of a piloting round. This was followed by an exploratory factor analysis in stage 1 exploring the underlying constructs of the EE questionnaire, reducing the items to 40. In a second stage these constructs were validated by calculating a confirmatory factor analysis, which was used to further reduce the items to 32. Overall, the whole procedure revealed eight factors with satisfactory reliability:

EE internalized	0.88
EE Gaming	0.92
EE Digital Creative	0.75
EE Nice Activities	0.71
EE Viewing	0.75
EE Social Interaction	0.89
EE music	0.57
EE Reading & Listening	0.74

Figure 11.7: Constructs of Sundqvist & Uztosun (under review) with reliability value

Comparing the categories with the constructs used for this study reveals that both studies have found similar components regarding English music, English viewing and English gaming. Moreover, both studies have another factor that is concerned with digital EE activities. However, while this study focuses on the use of apps and search engines in the construct English online, Sundqvist and Uztosun (under review) have included a broader variety of different digital activities, revealing a construct related to digital creative English use, which has not been considered in this study. Moreover, Sundqvist and Uztosun (under review) have one construct for *EE Social interaction*, which combines instances of spoken and written communication. Consequently, this would compare to the factors *Communication-oriented reading* and *Written communication* in this study as well as instances of *Speaking in English*. However, the construct *Speaking in English* also contains items similar to Sundqvist and Uztosun's (under review) component *EE internalized*. Furthermore, Sundqvist and

Uztosun (under review) have combined EE reading and listening activities in one construct, which is reflected in several constructs in this study. Therefore, the genre orientation of the constructs revealed is not as visible in Sundqvist and Uztosuns' (under review) findings, apart from reading and writing activities related to social interaction. As a result, future studies might want to further investigate the constructs underlying EE exposure, which will help to further interpret and validate given constructs.

Examining the amount of EE exposure of participants in this study it can be stated that participants showed a high involvement in EE activities with 97.4% of the participants performing at least one activity *nearly always* in English and participants on average engaging in ten EE activities *nearly always* and in 19 *often*. This stayed stable over the time of investigation with none of the computational variables changing significantly. Nonetheless, as in prior research studies (e.g. Olsson & Sylvén, 2015; Schwarz, 2020; Sylvén & Sundqvist, 2012) a large variety regarding amount of EE involvement can be reported, highlighting the extremely individual nature of EE. In this context, it has to be acknowledged that this research study cannot provide exact estimates of EE involvement in regards of time spent on various activities. This is due to the fact that the EE questionnaire developed was solely based on Likert scale items, which did not ask for exact time spent on an activity. Moreover, the labelling of the Likert scale items from *nearly always* to *nearly never* makes interpretation of the answers harder than necessary in retrospective. Therefore, future research should ensure more accurate labelling of Likert scale items and should strive for items giving estimates of time spent on EE activities. Consequently, other research instruments, such as EE diaries already in use in other studies (e.g. Olsson, 2011; Olsson, 2016; Sundqvist, 2009; Sundqvist & Wikström, 2015; Sylvén, 2006; Sylvén, 2019a) could be used as an additional tool to get a better insight into participants' EE activities.

While this research gap on exact time periods spent on EE activities by tertiary level students remains open, this study hopes to contribute to the field by analysing the type of EE exposure reported in comparison to prior literature, as well as the influence of biographical background variables on EE involvement. These relational findings reveal several similarities and differences with prior studies.

Starting with the similarities this study reveals a similar popularity of the EE activities listening to English music, watching English video material and using the Internet in English as earlier studies at tertiary level (e.g. Avello et al. 2019; González-Fernández, 2021; Kusyk, 2017; Liu, 2014; Liu, Lin & Wiley, 2016; Trinder, 2015) and other educational levels (e.g. De Wilde & Eyckmans, 2017; Lefever, 2010; Kuppens, 2010; Persson & Prins, 2012; Sundqvist, 2022). While at the first data collection point music was the most popular EE activity, consuming English Youtube videos was able to overtake the first place after a year, closely followed by watching English films and series.

Consequently, a focus on audio-visual EE activities on screen can be reported, with the dominance of the two media varying according to the use of English subtitles. While several studies (e.g. Montero Perez et al., 2014; Peters, 2019; Peters, Heynen & Puimège, 2016; Winke, Gass & Sydorenko, 2010) have revealed a positive impact of subtitles on language uptake this study has not pursued such an investigative goal. Instead participants' use and reasons for using subtitles have been examined a little closer. This has shown that participants either use subtitles rather frequently or very scarcely, with only few participants reporting to use them sometimes. Since this phenomenon has been consistently observed throughout this study it can be assumed that the use of subtitles appears to be perceived as an either-or matter by participants. This is reinforced by the fact that some students perceive subtitles as distracting, drawing their active attention towards the written word, while others report to consume them rather passively. Expanding more on the use of subtitles has further revealed several reasons for using them, namely movie-related parameters, such as accent of actors or pace of speech, use of unknown movie languages, conscious wish for language learning, or physical surroundings making subtitle use necessary. While all of these reasons were named at both times of investigation, movie-related parameters became more dominant over time, which might be partially explained by the increasing popularity of streaming platforms such as Netflix, which have integrated subtitles.

Zooming in into the use of the internet reveals a growing popularity of social media activities also reported in other contexts (e.g. Inozu, Sahinkarakas & Yumru, 2010; Kaur, 2015; Moncrief, 2011; Mukundan, Khojasteh & Pearson, 2009; Trinder, 2015). However, while Kusyk (2017, p. 83) and Arnbjörnsdóttir (2017, p. 49) state that their participants used online communication extensively both receptively and productively, this research study can only confirm the comparative popularity of reading social media messages, posts and status messages. Overall, findings regarding engagement in receptive and productive EE activities align with prior research studies (e.g. Arnbjörnsdóttir, 2017; Chusanachoti, 2009; Hyland, 2004; Kusyk, 2017; Lai & Zheng, 2018; Littlewood & Liu, 1996; Orhon, 2018; Pickard, 1996) reporting larger receptive involvement. Consequently, also participants of this study appear to be involved in EE writing activities, including written genres associated with social media platforms, as little as participants of other studies at tertiary level (e.g. Ekşi & Aydın, 2013; Ferdous, 2013; Kaur, 2015; Knight, 2007; Orhon, 2018). However, when comparing written communication, such as writing social media messages and status messages to other types of writing, such as genres associated with Creative writing, it still is the most popular written genre participants engage in.

Comparing participants' reading practices in English to other studies (e.g. Chan, Spratt & Humphreys, 2002; González-Fernández, 2021; Hyland, 2004; Inozu, Sahinkarakas & Yumru, 2010; Lamb, 2002; Liu, 2014; Liu, Lin & Wiley, 2016; Mukundan, Khojasteh & Pearson, 2009; Moncrief, 2011) confirms the trend of tertiary level students' higher involvement in English

reading activities detected in comparison to other educational levels (e.g. Aniol, 2011; Barbee, 2013; Fallkvist, 2016; Nilsson, 2013; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén & Sundqvist, 2012; Yi, 2005). Moreover, while prior studies (e.g. Ferdous, 2013; Hyland, 2004; Inozu, Shinakarakas & Yumru, 2010; Liu, 2014; Liu, Lin & Wiley, 2016; Moncrief, 2011; Mukundan, Khojasteh & Pearson, 2009; Orhon, 2018; Pickard, 1996) appear to disagree on the popularity of different genres, this study rather conforms with Inozu, Shinakarakas and Yumru's (2010, p. 6) as well as Moncrief's (2011, p. 112) findings that genres, such as novels and articles, traditionally associated with print media, are relatively popular amongst participants, with English books being often read by 80% of the participants. However, whereas the popularity of digital genres associated with communication-oriented reading, such as reading status messages and social media messages remains stable over time, participants of this study report to read English books significantly less over the time of investigation. This could be either explained by the emergence of new literacy practices as already proposed by other research studies (e.g. Arnbjörnsdóttir, 2017; Busby, 2020; Trinder, 2015; Warnby, 2022) or might be simply related to an increased reading load in their studies, keeping participants from reading more extensive and time consuming genres, such as novels in their spare time.

Parallel to an increased involvement of tertiary level students in EE reading practices, previous studies have reported a more varied engagement in English speaking activities. While some (e.g. Freeman, 1999; Knight, 2007; Lamb, 2002; Moncrief, 2011; Spratt, Humphreys & Chan, 2002) report an increased participation in English speaking situations at tertiary level, others (e.g. Broemer & Virkkunen-Fullenwider, 2011; Ferdous, 2013; Hyland, 2004; Kaur, 2015; Kusyk, 2017; Liu, Lin & Wiley, 2016; Pearson, 2004; Pickard, 1996; Sockett & Toffoli, 2012) still report a relatively low engagement due to either lack of opportunities, lack of initiative or cultural burdens. This study rather aligns with the latter findings, since only one English speaking activity is often performed by over 50% of the participants, which is thinking or talking to myself in English. While this might point towards a lack of speaking opportunities, as reported by Pickard (1996, p. 4), a majority of students (66%) report to have opportunities to speak English in their spare time, while only 34% report to lack this option.

11.6 The effect of biographical variables on EE and its stability over time

The difference of availability of speaking opportunities in their spare time has also shown to be a surprisingly strong and stable influential factor on participants' EE behaviour. Students who reported to have the option to speak English in their spare time used more English online and were more involved in Communication-oriented reading activities, as well as Creative writing and Written Communication activities. These differences remained constant over the time of investigation, demonstrating a certain degree of stability. Therefore, while prior studies have already revealed differences in EE behaviour based on

gender (Hannibal Jensen, 2017) or age (Hannibal Jensen, 2017) or different educational backgrounds (Lancaster, 2018), this variable offers new insights into the far-reaching effects limited oral communication opportunities can have on other aspects of EE.

Extending the comparison of results regarding the influence of biographical background on EE reveals even more similarities with findings of previous studies. As reported by Hannibal Jensen (2017) at primary school level and several studies at secondary school level (e.g. Grau, 2009; Greenberg et al., 2010; Olsson, 2011; Olsson & Sylvén, 2015; Peters, 2018; Sundqvist, 2009; Sundqvist & Sylvén, 2014; Sylvén, 2006; Sylvén & Sundqvist, 2012; Tran, 2017) this study likewise found gender differences in regard to playing computer games, with boys being more involved than girls. Moreover, similar to other studies girls appear to engage more in several other EE activities than boys. In agreement with Yi (2005), Sundqvist (2009), Sylvén & Sundqvist (2012) and Sundqvist & Sylvén (2014) amongst many others, female participants reported to be significantly more involved in Fiction-oriented reading activities consistently over the time of investigation, including classical novels and books. Similarly female participants of this study displayed a consistently higher involvement in listening to English music over the time of investigation than male participants, which aligns with Sylvén's (2006) and Schwarz's (2020) findings. However, fewer prior studies have found a gender difference regarding the consumption of English music, with girls consuming more English music than boys. Nonetheless, this difference remained stable over the whole time of investigation, similar to girls enhanced consumption of Fiction-oriented texts. Consequently, this trend might be specific for the observed context and might not be generalisable for other contexts.

Moving on to the three remaining potential influencing variables on EE behaviour shows that the effect of differences in age of onset and educational degree found at the beginning of the investigation ceases over time for all EE activities. While after one semester participants with a later age of onset used less English gaming, written communication and English speaking in their spare time across all groups, this effect was limited to specific target groups after a year of investigation. The same phenomenon was observed for the variable educational degree. Consequently, these findings highlight the importance of examining the development of EE over time as well as the persistence of effects of biographical variables.

This is additionally highlighted when looking at the influence of experience abroad, which showed to have a significant effect on written communication and English speaking involvement at the first data collection point. However, after one semester no differences between participants having been abroad and participants without experience abroad could be found anymore.

In summary the influence of the biographical variables gender and spare time speaking opportunities appears to be relatively stable over the time of investigation, while the effect

of age of onset and level of education decreases over time. The only fully volatile variable was experience abroad. Therefore, future EE studies should strive to investigate this concept longitudinally to capture changes in effects and behaviour. This will avoid jumping to conclusions which might not show prevalence over a longer period of time and will allow deeper insights into the concept of EE.

11.7 The effect of EE on academic verb knowledge

While Kremmel (2020) outlines that the usual way to investigate the influence of an independent on a dependent variable is to compare performance of an experimental and a control group, this procedure would have been difficult in the context of EE, since it would have been a challenge to find a similar sized control group which is not exposed to English in their spare time at all over one academic year. Therefore, no control group has been used in this study design. However, similar to studies such as Peters (2019) and Peters, Heynen and Puimège (2016), which are focusing exclusively on the impact of a particular EE activity, using an experimental group involved in the activity and a control group not involved in the activity, future studies might want to adopt an similar experimental study design to reveal more in-depth information about particular EE activities, which have shown to have a positive impact in large scale studies like this one.

Comparing the finding of this study to earlier research reveals that similar to the latter (Lee, 2019a; Li & Hafner, 2021; Peters et al., 2019; Sundqvist, 2009; Verspoor, De Bot & Van Rein, 2011) an influence of EE exposure on all receptive word knowledge types could be detected. Moreover, a similar positive influence of engagement in particular EE activities has been identified for productive form-meaning link knowledge. Consequently, this study rather aligns with studies (Olsson, 2016; Sundqvist, 2009), which reported that EE engagement has a positive effect on both, receptive and productive vocabulary knowledge. In contrast, other studies like Olsson and Sylvén (2015), Lee (2019a) and Schwarz (2020) could detect no significant influence of the quantity of EE on vocabulary in text production or scores on productive vocabulary knowledge tests. Therefore, future studies might still want to shed light on these divergent results related to productive vocabulary knowledge, considering comparability of test instruments, varying participant profiles or different word types focused at.

Puimège and Peters (2019, p. 3) have particularly highlighted one of these research gaps by stating that “research has not yet addressed the question of which words are more likely to be picked up from EE”. Responding to this call this research study was geared towards a very specific group of words, namely academic verbs, which has been focused on by only few earlier studies (e.g. Busby, 2020; Olsson & Sylvén, 2015; Sundqvist, 2009; Warnby, 2022).

Starting with Sundqvist (2009, p. 151), investigating the effect of EE on words from different frequency bands, it has been found that EE appears to have less effect on less frequent words. This potentially implies that EE might also have no big impact on academic word knowledge. This is further supported by Busby (2020) and Olsson & Sylvén (2015), who have found that EE appears to have no significant impact on receptive form-meaning link knowledge or productive use of academic words in student texts.

Nonetheless, as visible in Table 11.2 this study has found a significant connection between particular EE activities and all four types of academic word knowledge investigated.

2nd data collection		3rd data collection
Form-meaning recognition	Fiction-oriented reading (Creative writing)*	Creative writing (Fiction-oriented reading)*
Form-meaning recall	Creative writing	Creative writing (Fiction-oriented reading)*
Collocation recognition	Creative writing (English speaking)*	English online Fiction-oriented reading
Multiple meaning recognition	Fiction-oriented reading (Creative writing)*	Fiction-oriented reading

Table 11.2: EE activities predicting four different word knowledge components of academic verbs

This finding aligns with only one very recent study by Warnby (2022, p. 133), who also reports significant effects of EE on form-meaning recognition knowledge of academic words. Another striking similarity is the fact that both studies found that reading appears to have a positive impact on form-meaning recognition knowledge of academic verbs. However, while this study narrows the reading type influencing form-meaning recognition knowledge to Fiction-oriented reading, Warnby (2022) uses computational variables, out of which reading comprises fiction, non-fiction and news. Consequently, the overlap between the two studies regarding the positive impact on reading on form-meaning recognition knowledge has to be interpreted with caution, since Warnby's (2022, p. 131) computational variables combines Fiction-oriented reading with other variables having no significant impact in this study.

Moreover, Warnby (2022, p. 133) reports that also other EE activities, such as watching video material without and with English or Swedish subtitles, significantly predicts form-meaning recognition knowledge of academic words, which cannot be confirmed in this study. Instead a significant impact of Creative writing activities, such as writing English stories, comics, blogs or fanfiction on form-meaning recognition knowledge has been found, which is a category of EE activities not included in Warnby's (2022) study. What is particularly interesting regarding the balance between Fiction-oriented reading and Creative writing as predictors for form-meaning recognition knowledge is that their dominance appears to alternate over time. While at the 2nd data collection point Creative writing is only a significant predictor of form-meaning recognition knowledge when it is examined in isolation, it loses its predictive power when being combined with Fiction-oriented reading.

The exact reverse picture can be seen at the 3rd data collection point.⁹⁵ Consequently, while the types of predictors, namely Fiction-oriented reading and Creative writing, stay stable over time, their individual predictive roles appear to alternate.

Zooming in on the other three word knowledge components researched reveals another research gap waiting to be filled. While prior studies, such as Lee (2019a), Li and Hafner (2022) or Warnby (2022) have focused mainly on the influence of EE on receptive and productive vocabulary size, only one study to the author's knowledge (González-Fernández, 2021) has also extended this scope to further word knowledge components.

Looking at the predictors of form-meaning recall, collocation recognition and multiple meaning recognition knowledge reveals that at least Fiction-oriented reading or Creative writing appear to be predictors for all word knowledge aspects investigated across time. While knowledge of productive word knowledge components, such as form-meaning recall in this study, is rather predicted by similarly productive EE activities, such as Creative writing, receptive word knowledge components are explained by a mix of both, with Fiction-oriented reading being the most prominent predictor. This consistency of predictors across word knowledge components has also been observed by González-Fernández (2021), reporting that watching TV and reading functioned as positive predictors across several word knowledge components. However, it has to be considered that González-Fernández's (2021) study is not geared towards multiple component knowledge of academic words but instead focuses on knowledge of general words. Consequently, differences, such as the omission of watching TV as a predictor in this study might be partially explained by these different foci. This is further supported by the fact that multiple meaning knowledge, potentially containing also more specific secondary meanings, is only predicted by reading, similar to multiple meaning recognition in this study. Moreover, while reading appears as a significant predictor in both studies, Creative writing has not been incorporated in González-Fernández's (2021) study, which included only relatively broad EE categories, such as watching TV, reading, listening to music and engaging in social media or gaming.

Considering that Creative writing activities, such as writing stories, ($MD=2$, $IQR=2$), fanfiction ($MD=1$, $IQR=0$), comics ($MD=1$, $IQR=0$) and blogs ($MD=1$, $IQR=1$), are not involved in heavily on average, it has to be considered that the impact of Creative writing on form-meaning recall knowledge could be only due to more engagement of students with high scores on the Form-meaning Recall Test of Academic Verbs. This option was ruled out by calculating ANOVAs, which revealed no significant difference regarding the engagement of different proficiency groups concerning form-meaning recall knowledge (English stories $H(2)=1.248$, $p=0.536$; English fanfiction $H(2)=1.655$, $p=0.437$; English comics $H(2)=0.271$, $p=0.873$; English blogs $H(2)=1.155$, $p=0.561$). Consequently, this rather scarce involvement in Creative writing

⁹⁵ This has been visualized with an * in Table 11.2.

activities still appears to significantly impact academic form-meaning recall knowledge at both points in time.

Focusing on the longitudinal information about EE predictors provided by this study reveals, that for form-meaning recognition, recall and multiple meaning recognition knowledge predictors appear to be relatively stable. While form-meaning recognition knowledge is consistently explained by Fiction-oriented reading and Creative writing activities, with alternating power, form-meaning recall is explained by Creative writing at both points in time. However, at the third data collection point Fiction-oriented reading, while not being a significant predictor next to Creative writing anymore, adds some more explanatory power to the model. The reverse picture is visible for multiple meaning recognition, which is explained mostly by involvement in Fiction-oriented reading but has Creative writing as additional possibility to further increase explanatory power. The only word knowledge component which is influenced by a variety of different word knowledge components over time is collocation recognition knowledge of academic verbs. While collocation recognition knowledge is mostly predicted by English speaking activities⁹⁶, in combination with Creative writing at the 2nd data collection, both English online and Fiction-oriented reading significantly predict this knowledge type at the 3rd data collection. A possible explanation could be that Creative writing as well as Fiction-oriented reading are classical sources of in-depth contact with collocations for students, while English speaking and English online have a momentary impact on their academic collocation recognition. This could be explained by the fact that the participants are at the beginning of their studies, talking to new colleagues and friends also in their spare time about topics related to their studies. Consequently, also spare time English conversation might incorporate academic collocations. Furthermore, the EE category English online, includes using English apps as well as search engines, which both might require the entry of word combinations. However, this are only first speculations, which need to be substantiated by further studies.

	2nd data collection	3rd data collection
Form-meaning recognition	$R^2=8.0\%$, $R^2_a=7.1\%$	$R^2=9.8\%$, $R^2_a=8.8\%$
Form-meaning recall	$R^2=5.5\%$, $R^2_a=4.5\%$	$R^2=13.8\%$, $R^2_a=12.0\%$
Collocation recognition	$R^2=9.9\%$, $R^2_a=8.1\%$	$R^2=9.8\%$, $R^2_a=8.0\%$
Multiple meaning recognition	$R^2=10.4\%$, $R^2_a=9.5\%$	$R^2=6.8\%$, $R^2_a=5.9$

Table 11.3: Predictive power of EE activities

⁹⁶ The component English speaking comprises the following items: speaking English on Skype or other internet services, speaking English on the phone, speaking English in face to face interactios and thinking or talking to oneself.

Taking a look at the explanatory power of the significant EE predictors on the different word knowledge components as visible in Table 11.3 reveals that the explained variability of the scores through EE is rather small.

At the same time, these rather low results appear to be plausible considering the complexity of the construct researched. As already foreshadowed by the SEM model vocabulary knowledge is a highly complex construct, whose word knowledge components are explained by several factors out of which EE contact is only one potential predictor variable.

Focusing on the longitudinal development of the predictive power of the EE predictors on the different types of word knowledge reveals the following picture :

Form-meaning recognition	2nd data collection	<	3rd data collection
Form-meaning recall		<	
Collocation recognition		=	
Multiple meaning recall		>	

Table 11.4: Developmental patterns of predictive power

While the explanatory power of engagement in Fiction-oriented reading and Creative writing appears to increase over the time of investigation for both receptive and productive form-meaning link knowledge, collocation recognition knowledge is predicted by EE activities to a similar extend at both points in time. The only word knowledge component for which EE loses explanatory power over time is multiple meaning recognition knowledge. Nevertheless, further studies will have to be conducted to provide more data about the development of the explanatory power of EE in regards to different word knowledge aspects and which potential explanations could stand behind these trends.

11.8 Implications for the educational context

As a considerable pedagogical potential is slumbering within the concept of EE this chapter tries to point at some pertinent insights related to the findings of this research study. At the same time, formulating implications for an educational context is a difficult matter, since EE is per definition non teacher-initiated English language contact which is happening outside of an educational context. Consequently, while this study amongst other research projects (González-Fernández, 2021; Warnby, 2022), has revealed that EE engagement is supportive for academic vocabulary development, the direct transmission of EE activities, such as creative writing or fictional reading into class might fail to achieve the same positive effect.

Therefore, a more indirect approach to apply the results in educational practice appears to be necessary. As a first step educators should be made aware of the potential that is slumbering within EE contact. This might require more detailed information about the

different dimensions of EE and how EE can be defined. For this purpose the EE model introduced in chapter 5.1 might be used as a reflection tool for teachers (see Figure 5.7):

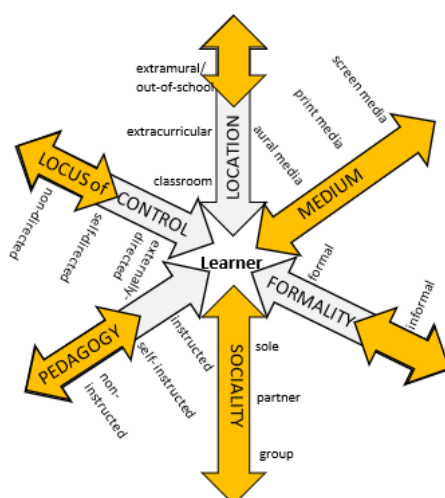


Figure 5.7: Model of Extramural English

In a next step, educators might familiarize their students with the concept of Extramural English, highlighting the positive impact that certain EE activities can have on their language development. More specifically, lecturers at tertiary level might highlight that engagement in EE does not only foster students' general language development but that enhanced attention to creative writing and reading fictional texts particularly improves their knowledge of academic words.

Additional to general awareness raising, educators might inform students about different options to engage in EE. In a university context this might include broadening students' recognition of different text genres, since different fictional texts appear to improve several aspects of academic word knowledge. While most students will consider reading an English novel, fewer will be familiar with fanfiction or might not have thought about reading an English comic. A similar approach might be taken for creative writing activities, offering students ideas for innovative text productions, including different genres like blogs, forum entries, comics, mangas or fanfiction. In addition to pointing out different text types that could be produced, educators might offer prompts that should spark an idea and get students writing. However, prompts should never be obligatory but should be an offer for students who question their own creativity. This might counteract students' reluctance to involve in creative writing activities and should help them to use creative writing tasks to enhance their academic word knowledge.

As fictional reading and creative writing activities have the broadest impact on knowledge of academic words, it is proposed that EE activities requiring intense active engagement with a language seem to be beneficial for academic vocabulary development across word knowledge aspects. Consequently, students can be reminded of the fact that

communication is one of the main goals in language learning⁹⁷ and might be motivated to engage in EE activities which require them to work actively with the language.

This goes hand in hand with the idea that future teaching and research should focus more on the transition between receptive and productive word knowledge. As outlined by González-Fernández and Schmitt (2019, p. 21) many vocabulary learning tasks typically focus on receptive presentation of a word's meaning, while neglecting necessary repetition and other ways to productive mastery. Hence, future teaching materials and tasks might want to integrate practice of recall knowledge more intensively.

12 Conclusion

To sum up, building on past research (e.g. González-Fernández & Schmitt, 2019; Gyllstad, 2007; Milton & Fitzpatrick, 2014; Nation, 2013), which has established vocabulary knowledge as a highly complex construct, comprising both vocabulary size and vocabulary depth, this study is aiming at a close investigation of the knowledge of academic verbs displayed by undergraduate students majoring in English. Following Nation's (2013) taxonomy of word knowledge, the focus lies on 147 Viennese English major students' form-meaning recognition, form-meaning recall, multiple-meaning recognition and collocation recognition knowledge of academic verbs across frequency levels. For this purpose four tests, one for each word knowledge component, were developed, containing the same academic verbs. By following the students for one academic year and testing them at three points in time, light can be shed on the longitudinal development of these word knowledge components. In a nutshell it can be stated that all word knowledge aspects researched displayed a significant development over time. As visible by the effect sizes measured, form-meaning recall knowledge increased the most, followed by multiple meaning recognition knowledge, collocation recognition knowledge and finally form-meaning recognition knowledge. These differences in effect size rather suggest a non-linear developmental curve, which reaches a plateau at higher proficiency levels. Additionally, word knowledge components in this study are not only researched as individual, detached categories but are viewed holistically, examining the relationships between them. Using SEM modelling additional to correlational analysis a fundamental distinction between recognition and recall knowledge, similar to González-Fernández & Schmitt (2019) was revealed. However, different word knowledge components appear to be explained by the construct of vocabulary knowledge to a different degree. While form-meaning link knowledge, classically associated with vocabulary size, is explained the most by an overarching construct of

⁹⁷ This can be done by including guided in class reflections about why students want to learn a language. This can also be very useful in teacher education to remind them that their own students will most likely strive for a similar goal and can be a good transition to introduce or remind of the principles of Communicative Language Teaching (e.g. Barcroft, 2015; Hedge, 2000; Richards, 2006).

vocabulary knowledge, both collocation recognition and multiple meaning recognition knowledge revealed smaller factor loadings. Consequently, it seems as if word knowledge aspects traditionally associated with vocabulary depth (e.g. Qian, 2002, p. 515) are explained by additional factors.

One of these additional factors influencing academic vocabulary knowledge is the second construct researched in this study, namely Extramural English. Extramural English, being a term coined by Sundqvist (2009) has been defined as non teacher-initiated English language contact located outside of an educational context (Sundqvist & Sylvén, 2016, p. 11). Furthermore, the language engagement can be both intentional or incidental covering both active or passive activities (Sundqvist & Sylvén, 2016, p. 11). Using a self-designed questionnaire, the survey elicits students' non-teacher-initiated activities involving English done outside their educational context (Sundqvist & Sylvén, 2016; Sylvén, 2019a). Similar to their academic vocabulary knowledge, students reported their Extramural English behaviour over one academic year, filling out the questionnaire after each semester. Overall, English major students in an Austrian setting displayed a high EE involvement, showing most interest in listening to English music, consuming English video material and reading practices in English similar to prior studies at tertiary level (e.g. Avello et al. 2019; González-Fernández, 2021; Liu, Lin & Wiley, 2016; Trinder, 2015). Moreover, while Extramural English behaviour at tertiary level has already been researched in other settings (e.g. Busby, 2020; Lee & Sylvén, 2021; Sylvén & Sundqvist, 2012; Sundqvist & Sylvén, 2014), this study additionally presents a rare longitudinal perspective on Extramural English engagement, revealing a remarkably stable involvement in EE activities and differences related to biographical background variables, such as gender or existence of spare time speaking opportunities over time. For instance, frequently reported differences regarding boys participating more in playing English computer games (e.g. Peters, 2018; Sylvén, 2006; Sylvén & Sundqvist, 2012) and girls involving more in reading fiction-oriented genres (e.g. Sylvén & Sundqvist, 2012; Yi, 2005) or listening to English music (e.g. Schwarz, 2020; Sylvén, 2006) were found to be stable over one academic year.

By linking Extramural English and vocabulary knowledge, this study builds on a pool of recent research (e.g. Cole & Vanderplank, 2016; Lee, 2019a; Lee & Dressman, 2018) highlighting the importance of Extramural English for vocabulary knowledge. However, a new angle is presented by zooming in on academic, rather than general, vocabulary knowledge and even more specifically on academic verb knowledge. Using single and multiple linear regression models, the influence of Extramural English activities on the different aspects of academic verb knowledge is investigated, revealing that particular types of Extramural English activities are significant predictors for academic verb knowledge. Especially the consistency of Fiction-oriented reading and Creative writing as predictors across the word knowledge components researched is interesting, leading to the proposal to raise students' awareness of EE in general and different, less widely practiced but

successful EE activities, such as reading fanfiction or composing comics in English, in particular.

While this study revealed interesting findings, future studies should strive for a higher participant number, which was constrained in this study by the specific target group chosen and administrative surroundings in relation to the data collection. Moreover, this study has developed a unique test battery, which might undergo even more validation steps as were possible within the scope of this research. In general, prospective studies should seek to take creative approaches to testing several word knowledge components with test batteries, including solutions for guessing behaviour and constraints of time. These results might then be used to further examine different models for vocabulary knowledge and how it is connected to other skills and factors. In more detail the development of different word knowledge components building on different theoretical models, such as Bordag, Gor and Opitz's (2022) ontogenesis model might be examined.

Regarding the field of Extramural English more studies should strive to examine the concept of EE closer through using advanced statistical methods, such as factor analysis or SEM. Unfortunately, this study was also only able to take a first step into this direction due to the limited participant number, which did not allow the calculation of a factor analysis for the whole construct. Moreover, EE was measured solely via the use of a self-report questionnaire and did not include any additional measurement instruments. Consequently, future studies might want to include a larger scope of different research instruments existent and might take new creative approaches at how to measure EE without relying on participant self-report. Lastly, forthcoming studies might also want to add to the relatively small pool of longitudinal EE research, allowing insights into the stability of EE behaviour and its influence on different aspects of language proficiency, such as vocabulary knowledge.

References

- Abou-Khalil, Victoria; Helou, Samar; Flanagan, Brendan; Chen, Mei-Rong Alice; Ogata, Hiroaki. 2019. "Learning isolated polysemous words: identifying the intended meaning of language learners in informal ubiquitous language learning environments". *Smart Learning Environments* 6(13), 1-18.
- Ackermann, Kirsten; Chen, Yu-Hua. 2013. "Developing the Academic Collocation List (ACL): A corpus driven and expert-judged approach". *Journal of English for Academic Purposes* 12, 235-247.
- Adelian, Maryam; Nemati, Azadeh; Fumani, Mohammad R. 2015. "The effect of Iranian advanced EFL learners' knowledge of collocation on their writing ability." *Theory and Practice in Language Studies* 5(5), 974-980.
- Aijmer, Karin. 2009. "So er I just sort of I dunno I think it's just because..: A corpus study of "I don't know" and "dunno" in learner spoken English". In Jucker, Andreas H.; Schreier, Daniel; Hundt, Marianne (eds.). *Corpora: Pragmatics and discourse*. Amsterdam: Rodopi, 151–166. ISBN: 978-90-420-2592-9.
- Aitchison, Jean. 1987. *Words in the mind: An introduction to the mental lexicon*. Oxford: Blackwell.
- Alnamer, Abdul S. 2017. "On the awareness of English polysemous words by Arabic-speaking EFL learners". *Advances in language and literary studies* 8(2), 112. DOI: 10.7575/aiac.all.v.8n.2p.112
- Alqraini, Faisl M.; Paul, Peter V. 2020. "The effects of a vocabulary intervention on teaching multiple-meaning words to students who are d/deaf and hard of hearing". *Journal of Deaf Studies and Deaf Education*, 469-489.
- Alsakran, Rayed A. 2011. "The productive and receptive knowledge of collocations by advanced Arabic-speaking ESL/EFL learners." PhD thesis, Colorado State University.
- Altenberg, Bengt. 1998. "On the phraseology of spoken English: The evidence of recurrent word-combinations". In Cowie, Anthony (ed.). *Phraseology: Theory, analysis, and application*. Oxford: Clarendon Press, 101–122.
- Anderson, John. 1980. "The lexical difficulties of English medical discourse for Egyptian students". *English for Specific Purposes* 37, 4.
- Anderson, Prue; Morgan, George. 2008. *Developing tests and questionnaires for a national assessment of educational achievement*. Washington: The World Bank.
- Anderson, Richard; Freebody, Peter. 1981. "Vocabulary knowledge". In Guthrie, John T. (ed.). *Comprehension and teaching: Research reviews*. Newark, DE: International Reading Association, 77–117.

- Aniol, Magdalena. 2011. "New media and new literacies: Mapping extracurricular English language competences of Polish and Norwegian adolescence". In Kaczmarek, Maria (ed.). *Health and well-being in adolescence: Part two Media*. Poznań: Bogucki Wydawnictwo Naukowe, 101-124.
- Anthony, Laurence. 2022. *AntWordProfiler* (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from <https://www.laurenceanthony.net/software>
- Anthony, Laurence; Nation, Paul I.S. 2017. *Picture Vocabulary Size Test (PVST)*. Retrieved from <https://www.laurenceanthony.net/software.html>
- Ardasheva, Yuliya; Tretter, Thomas R. 2017. "Developing science-specific, technical vocabulary of high school newcomer English learners". *International Journal of Bilingual Education And Bilingualism* 20, 252-271.
- Arnbjörnsdóttir, Birna. 2017. "English exposure, proficiency and use in Iceland". In Arnbjörnsdóttir, Birna; Ingvarsdóttir, Hafdís (eds.). *Language development across the life span: The impact of English on education and work in Iceland*. Cham: Springer, 35-55.
- Arnold, Nike. 2009. "Online extensive reading for advanced foreign language learners: An evaluation study". *Foreign Language Annals* 42(2), 340-366.
- Arnon, Inbal; Priva, Uriel C. 2013. "More than words: Frequency effects for multi-word phrases". *Journal of Memory and Language* 62(1), 67-82.
- Asaad, Hani Q.M.; Shabdin, Ahmad A. 2021. "The predictive role of morphological awareness and productive vocabulary knowledge in L2 postgraduate students' academic writing". *Eurasian Journal of Applied Linguistics* 7(1), 24-44.
- Avello, Maria; Camacho-Miñano, María-del-Mar; Urquía-Grande, Elena; del Campo, Cristina. 2019. "Do you use English in your daily life?": Undergraduate students' perceptions of their extramural use of English". *Journal of Teaching in International Business* 30(1), 77-94.
- Aviad-Levitzky, Tami; Laufer, Batia; Goldstein, Zahava. 2019. "The New Computer Adaptive Test of Size and Strength (CATSS): Development and Validation". *Language Assessment Quarterly* 16(3), 345-368. doi:10.1177/0265532209340194
- Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, Lyle F. 2004. *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
- Barbee, Matthew. 2013. "Extracurricular L2 input in a Japanese EFL context: Exposure, attitudes, and motivation". *Second Language Studies* 32(1), 1-58.

- Barcrof, Joe. *Vocabulary in communicative language teaching*. Routledge. ISBN: 9781138931763
- Barfield, Andrew W. 2006. "An exploration of second language collocation knowledge and development." PhD thesis, Swansea University.
- Batluk, Liilia. 2015. "Does listening to English songs motivate students to expand their Extramural English?: A case study of Swedish upper-secondary school students' perceptions of possible impact of listening to vocal music on their Extramural English". BA thesis, Linné Universitetet.
- Bear, Donald R.; Invernizzi, Marcia; Templeton, Shane; Johnston Francine. 2012. *Words their way: Word study for phonics, vocabulary, and spelling instruction*. (5th edition). Boston: Allyn & Bacon.
- Beck, Isabel L.; McKeown Margret G.; Kucan, Linda. 2002. *Bringing words to life: Robust vocabulary instruction*. New York: Guilford.
- Becka, John V. 1972. "The lexical composition of specialised texts and its quantitative aspect". *Prague Studies in Mathematical Linguistics* 4, 47-64.
- Begagić, Mirna. 2014. "English language students' productive and receptive knowledge of collocations". *ExELL* 2(1), 46-67.
- Beglar, David. 2010. "A Rasch-based validation of the Vocabulary Size Test". *Language Testing* 27(1), 101–118. doi:10.1177/0265532209340194
- Bengtsson, Andreas. 2014. "Watching video or studying? An investigation of the extramural activities and Japanese language proficiency of foreign language learners of Japanese". MA thesis, Stockholm University.
- Benson, Phil. 2001. *Teaching and researching autonomy in language learning*. Harlow: Pearson.
- Benson, Phil. 2011. "Language learning and teaching beyond the classroom: An introduction to the field". In Benson, Phil; Reinders, Hayo (eds.). *Beyond the language classroom*. Houndsmill, Basingstoke: Palgrave Macmillan, 6–16.
- Benson, Morton; Benson, Evelyn; Ilson, Richard. 1997. *The BBI dictionary of English word combinations*. Amsterdam: John Benjamins.
- Bergström, Denise; Norberg, Cathrine; Nordlund, Marie. 2021. "Words are picked up along the way": Swedish EFL teachers' conceptualizations of vocabulary knowledge and learning." *Language Awareness*, 1-17. DOI: 10.1080/09658416.2021.1893326
- Bernier-Colborne, Gabriel; Drouin, Patrick. 2014. "Creating a test corpus for term extractors through term annotation". *Terminology* 20(1), 50-73.

- Berns, Margie; De Bot, Kees; Hasebrink, Uwe (eds.). 2007. *In the presence of English: Media and European youth*. New York: Springer.
- Biber, Douglas. 2006. "Lexical bundles in university teaching and textbooks". In Biber, Douglas (ed.). *University language: A corpus based study of spoken and written registers*. Amsterdam: John Benjamins, 133-175.
- Biemiller, Andrew; Slonim Naomi. 2001. "Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition". *Journal of Educational Psychology* 93(3), 498-520.
- Binder, Katherine S.; Gilbert Cote, Nicole; Lee, Cheryl; Bessette, Emily; Vu, Huong. 2017. "Beyond breadth: The contributions of vocabulary depth to reading comprehension among skilled readers". *Journal of Research in Reading* 40(3), 333-343.
- Block, David. 2000. "Problematizing interview data: Voices in the mind's machine?". *TESOL Quarterly* 34(4), 757-763.
- Bonk, William J. 1996. "Testing learner's knowledge of collocations. " MA thesis, University of Hawaii at Manoa.
- Bonnet, Gérard (ed.). 2002. The assessment of pupils' skills in English in eight European countries. <https://www.educacionyfp.gob.es/inee/dam/jcr:d426c134-5c58-44f0-8641-3b5e4354ed37/habilidadesingles2002-1.pdf> (13 Nov. 2020).
- Booton, Sophie A.; Wonnacott, Elizabeth; Hodgkiss, Alex; Mathers, Sanders; Murphy, Victoria A. 2021. "Children's knowledge of multiple word meanings: Which factors count and for whom?". *Applied Linguistics* 43(2), 293–315.
- Bordag, Densia; Gor, Kira; Opitz, Andreas. 2022. "Ontogenesis Model of the L2 lexical representation". *Bilingualism: Language and Cognition* 25, 185–201.
<https://doi.org/10.1017/S1366728921000250>
- Brevik, Lisbeth M. 2016. "The gaming outliers: Does out-of-school gaming improve boys' reading skills in English as a second language?". In Elstad, Eyvind (ed.). *Educational technology and polycontextual bridging*. Rotterdam: Sense Publishers, 39–61. ISBN: 978-94-6300-645-3
- Brezina, Vaclav; Gablasova, Dana. 2015. "Is there a core general vocabulary?: Introducing the New General Service List". *Applied Linguistics* 36(1), 1-22.
- Brieger, Nick; Phol, Anton. 2002. *Technical English: Vocabulary and grammar*. Andover: Heinle Cengage Learning.
- Broemer, Marlene; Virkkunen-Fullenwider, Anu. 2011. "Students on their own: Listening activities outside the classroom without teacher supervision". Pitkänen, Kari K.; Jokinen, Jaana; Karjalainen, Sinikka; Karlsson, Leena; Lehtonen, Tuula; Matilainen,

- Mirjami; Niedling, Christian; Siddall, Roy (eds.). *Language Center Publications 2*. Helsinki: Helsingin yliopiston kielikeskus, 119-134. ISBN978-952-10-6901-7
- Brown, Fredrick, G. 1983. *Principles of educational and psychological testing*. New York: Holt, Rinehart and Winston.
- Brown, James A.; Fishco, Vivian V.; Hanna, Gerald S. 1993. *Nelson-Denny Reading Test: Manual for scoring and interpretation, Forms G & H*. Rolling Meadows, IL: Riverside Publishing.
- Browne, Charles. 2013. "The New General Service List: Celebrating 60 years of vocabulary learning". *The Language Teacher JALT2013 Special Issue* 4(37), 13-16.
- Browne, Charles; Culligan, Brent; Phillips, Joseph. 2013. "A new academic word list". Retrieved from www.newgeneralservicelist.org/nawl-new-academic-word-list (26.03.2021).
- Brunner, Edgar; Konietzschke, Frank; Bathke, Arne C.; Pauly, Markus. 2018. "Ranks and pseudo-ranks: Paradoxical results of rank tests". *arXiv Vanity*, 1-19. <https://doi.org/10.48550/arXiv.1802.05650>
- Brysbaert, Marc; Stevens, Michael; Mander, Paweł; Keuleers, Emmanuel. 2016. "How many words do we know?: Practical estimates of vocabulary size dependent of word definition, the degree of language input and the participant's age". *Frontiers in Psychology* 7(7), 1–11.
- Brysbaert, Marc; Mander, Paweł; Keuleers, Emmanuel. 2018. "The word frequency effect in word processing: A review update." *Current Directions in Psychological Science* 21(1), 2-16.
- Bundesministerium Bildung, Wissenschaft und Forschung (2005) Volksschullehrplan. https://www.bmbwf.gv.at/Themen/schule/schulpraxis/lp/lp_vs.html
- Bunting, Leona; Lindström, Berner. 2013. "Framing English learning at the intersection of school and out-of-school practices". *Journal of International Scientific Publications: Language, Individual & Society* 7(1), 205–221.
- Busby, Nicole L. 2020. "Words from where?: Predictors of L2 English vocabulary among Norwegian university students". *ITL - International Journal of Applied Linguistics*, 171(1), 1-27. <https://doi.org/10.1075/itl.19018.bus>
- Cabau, Béatrice. 2009. "The irresistible rise and hegemony of a linguistic fortress: English teaching in Sweden". *International Multilingual Research Journal* 3(2), 134– 152. <https://doi.org/10.1080/19313150903073786>
- Cabot, Michel. 2016. "In or out of school?: Meaningful output with digital and non-digital artefacts within personal English learning ecologies". *Nordic Journal of Digital Literacy* 11(3), 165–184.

- Cabré, Maria. 1999. *Terminology: Theory, methods, and applications*. Amsterdam: John Benjamins.
- Cambridge University Press. (n.d.). In Cambridge English Dictionary. Retrieved April 3, 2020 from <https://dictionary.cambridge.org/>
- Cameron, Lynne. 2002. "Measuring vocabulary size in English as an additional language." *Language Teaching Research* 6(2), 145-173.
- Capel, Annette. 2012. "Completing the English Vocabulary Profile: C1 and C2 vocabulary". *English Profile Journal* 3, 1–14.
- Carlisle, Joanne F. 2000. "Awareness of the structure and meaning of morphologically complex words: Impact on reading". *Reading and Writing: An Interdisciplinary Journal* 12, 169-190.
- Carroll John B. ; Davies, Peter; Richman, Barry. 1971. *The American Heritage Word Frequency Book*. New York: American Heritage Publishing Co. Inc.
- Carter, Ronald. 1987. *Vocabulary: Applied linguistic perspectives*. London: Allen & Unwin.
- Catalán Jiménez, Rosa M.; Ruiz de Zarobe, Yolanda; Cenoz Iragui, Jasone. 2006. "Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language". *VIEWS* 15(3), 23-27.
- Chan, Victoria; Spratt, Mary; Humphreys, Gillian. 2002. "Autonomous language learning: Hong Kong tertiary students' attitudes and behaviours". *Evaluation & Research in Education* 16(1), 1–18.
- Chappelle, Carol. 1998. "Construct definition and validity inquiry in SLA research". In Bachman, Lyle F.; Cohen, Andrew D. (eds.). *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 32-70.
- Charles, Maggie. 2014. "English for academic purposes". In Paltridge, Brian; Starfield, Sue (eds.). *The Handbook of English for Specific Purposes*. Chichester: John Wiley & Sons, Ltd, 137-153.
- Chen, Chuntien; Truscott, John. 2010. "The effects of repetition and L1 lexicalisation on incidental vocabulary acquisition". *Applied Linguistics* 31, 693-713.
- Chen, Qi; Ge, Guang-chun. 2007. "A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs)". *English for Specific Purposes* 26, 502–514 .
- Cheng, Junyu; Matthews, Joshua. 2018. "The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading". *Language Testing* 55(1), 3-25.
- Christie, Agatha. 1982. *Elephants can remember*. London: Collins.

- Chui, Amy. 2006. "A study of the English vocabulary knowledge of university students in Hong Kong." *Asian Journal of English Language Teaching* 16, 1-23.
- Chung, Teresa M.; Nation, Paul I.S. 2003. "A corpus comparison approach for extracting terminology". *Terminology* 9, 221-245.
- Chung, Teresa M.; Nation Paul I.S. 2004. "Identifying technical vocabulary". *System* 32, 251-263.
- Church, Kenneth W.; Hanks, Patrick. 1990. "Word association norms, mutual information, and lexicography". *Computational Linguistics* 16, 22–29.
- Chusanachoti, Ruedeerath. 2009. "EFL learning through language activities outside the classroom: A case study of English education students in Thailand". PhD thesis, Michigan State University.
- Cobb, Tom. 2000. "One size fits all?: Francophone learners and English vocabulary tests". *The Canadian Modern Language Review* 57(2), 295-324.
- Cobb, Tom. VocabProfilers [computer program]. Accessed 25 December 2021 at <https://www.lex tutor.ca/cgi-bin/range/texts/index.pl>
- Cobb, Tom. What texts can I read if I know this many words?: Reading a text depends on knowing at least 95% of its lexical words, so. Retrieved from <https://www.lex tutor.ca/vp/comp/samples.html> (28.07.2022)
- Cobb, Tom; Horst, Marliese. 2004. "Is there room for an AWL in French?". In: Bogaards, Paul; Laufer, Batia (eds.). *Vocabulary in a second language: Selection, acquisition, and testing*. Amsterdam: John Benjamins, 15-38.
- Cohen, Andrew; Glasman, Hilary; Rosenbaum-Cohen, Phyllis R.; Ferrara, Jonathan; Fine Jonathan. 1988. "Reading English for specialized purposes: Discourse analysis and the use of standard informants". In Carrell, Patricia; Devine, Joanne; Eskey, David (eds.). *Interactive approaches to second language reading*. Cambridge: Cambridge University Press, 152-167.
- Cohen, Jacob; Cohen, Patricia; West, Stephen G.; Aiken, Leona S. 2003. *Applied multiple regression/correlation analysis for the behavioral sciences*. (3rd edition). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cole, Jason; Vanderplank, Robert. 2016. "Comparing autonomous and class-based learners in Brazil: Evidence for the present-day advantages of informal, out-of-class learning". *System* 61, 31–42.
- Conklin, Kathy; Schmitt, Norbert. 2008. "Formulaic sequences: Are they processed more quickly than non-formulaic sequences by native and nonnative speakers?". *Applied Linguistics*, 29(1), 72-89.

- Conklin, Kathy; Schmitt, Norbert. 2012. "Processing of formulaic language". *Annual Review of Applied Linguistics* 32, 45-61.
- Cowie, Anthony P. 1988. "Stable and creative aspects of vocabulary use". In Carter, Ronald; McCarthy, Michael (eds.). *Vocabulary and Language teaching*. London: Longman, 126-139.
- Cowie, Anthony P. 1991. "Multiword units in newspaper language". In Granger, Sylviane (ed.). *Perspectives on the English Lexicon: A tribute to Jaques van Roey*. CILL 17(1-3), 101-116.
- Cowie, Anthony P. 1994. "Phraseology". In Asher, Ron E. (ed.). *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon, 3168-3171.
- Cowie, Anthony P. (ed.). 1998a. *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press.
- Cowie, Anthony P. 1998b. "Introduction". In Cowie, Anthony P. (ed.). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press, 1-20.
- Cowie, Anthony P. 1998c. "Phraseological dictionaries: Some East-West comparisons". In Cowie, Anthony P. (ed.). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press, 209-228.
- Coxhead, Averil. 2000. "A new academic word list". *TESOL Quarterly* 34(2), 213-238. <https://doi.org/10.2307/3587951>
- Coxhead, Averil. 2018. *Vocabulary and English for specific purposes research: quantitative and qualitative perspectives*. Florence: Routledge.
- Coxhead, Averil. 2020. "Academic vocabulary". In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 97-110.
- Coxhead, Averil; Hirsh, David. 2007. "A pilot science word list for EAP". *Revue française de linguistique appliquée* 12(2), 65-78.
- Coxhead, Averil; Byrd, Patricia. 2012. "Collocations and academic word lists: The strong, the weak and the lonely". In Moskowich, Isabel; Crespo, Begona (eds.). *Encoding the Past, Decoding the Future: Corpora in the 21st Century*. Newcastle upon Tyne: Cambridge Scholars Publishing, 1-28.
- Coxhead, Averil; Nation, Paul; Sim, Dalice. 2015. "Measuring the vocabulary size of native speakers of English in New Zealand secondary schools." *NZ J Educ Stud* 50, 121-135.
- Coxhead, Averil; Demecheleer, Murielle; McLaughlin, Emma. 2016. "The technical vocabulary of carpentry: loads, lists and bearings." *THE TESOLANZ JOURNAL* 24, 38-71.

- Coxhead, Averil; Boutorwick, Thomas J. 2018. "Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science." *TESOL Quarterly* 52(3), 588-610. doi: 10.1002/tesq.450
- Coyle, Do; Meyer, Oliver. 2021. *Beyond CLIL pluriliteracies teaching for deeper learning*. Cambridge: Cambridge University Press.
- Cronbach, Lee J. 1942. "An analysis of techniques for diagnostic vocabulary testing". *Journal of Educational Research* 36, 206-217.
- Cronbach, Lee J. 1951. "Coefficient alpha and the internal structure of tests". *Psychometrika* 16(3), 292-334.
- Cronbach, Lee J. 1984. *Essentials of Psychological Testing*. New York: Harper and Row.
- Dabbagh, Ali; Enayat, Mostafa J. 2019. "The role of vocabulary breadth and depth in predicting second language descriptive writing performance". *The Language Learning Journal* 47(5), 575-590.
- Dale, Edgar. 1965. "Vocabulary measurement: Techniques and major findings". *Elementary English* 42(8), 895-901.
- Daller, Helmut; Milton, James; Treffers-Daller, Jeanine. 2007. *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.
- Dang, Thi N.Y.; Coxhead, Averil; Webb, Stuart. 2017. "The academic spoken word list". *Language Learning* 67(4), 959-997.
- Daniel, Isabell R.; Kremmel, Benjamin. 2020. "Test review: Current options in at-home language proficiency tests for making high-stakes decisions". *Language Testing* 37(4), 600-619.
- Davies, Alan; Brown, Annie; Elder, Cathie; Hill, Kathryn; Lumley, Tom; McNamara, Tim. 1999. *Studies in Language Testing 7: Dictionary of Language Testing*. University of Cambridge Local Examinations Syndicate. Cambridge: Cambridge University Press.
- Davies, Mark. 2008. The Corpus of Contemporary American English (COCA): 600 million words, 1990-present. Available online at <https://www.english-corpora.org/coca/>
- DeChazal, Edward. 2014. *English for Academic Purposes*. Oxford: Oxford University Press.
- Deep GmbH. (n.d.). Linguee. Retrieved April 3, 2020 from <https://www.linguee.com/>
- DeHaan, Jonathan; Reed, Michael W.; Kuwada, Katsuko. 2010. "The effect of interactivity with a music video game on second language vocabulary recall". *Language Learning & Technology* 14(2), 74-94.

- DeKeyser, Robert. 2003. "Implicit and explicit learning". In Doughty, Catherine J.; Long, Michael H. (eds.). *The handbook of second language acquisition*. Malden: Blackwell, 313-348.
- Dentisak, Dokchandra. 2019. "Exploring Thai university students' receptive and productive knowledge of collocations across four academic faculties". *Advances in Language and Literary Studies* 10(6), 115-123.
- DeWilde, Vanessa; Eyckmans, June. 2017. "Game on! Young learners' incidental language learning of English prior to instruction". *Studies in Second Language Learning and Teaching* 7(4), 673-694.
- Dörnyei, Zoltán. 2007. *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
- Drew, Paul; Holt, Elizabeth. 1998. "Figures of speech: Figurative expressions and the management of topic transition in conversation". *Language in Society* 27, 495-522.
- Duff, Dawna; Brydon, Melissa. 2020. "Estimates of individual differences in vocabulary size in English: How many words are needed to 'close the vocabulary gap'". *Journal of Research in Reading* 43(4), 454-481.
- Dunn, Douglas M.; Dunn, Lloyd M. 1959. *Peabody Picture Vocabulary Test*. Circle Pines, MN: AGS.
- Dunn, Douglas M.; Dunn, Lloyd M. 1981. *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN : AGS.
- Dunn, Douglas M. ; Dunn, Lloyd M. 1997. *Peabody Picture Vocabulary Test—3*. Circle Pines, MN: AGS.
- Dunn, Lloyd M.; Dunn, Douglas M. 2007. *Peabody Picture Vocabulary Test (4th ed.)*. Circle Pines, MN: American Guidance Service.
- Durrant, Philip. 2009. "Investigating the viability of a collocation list for students of English for academic purposes". *English for Specific Purposes* 28, 157-169.
- Ebel, Robert L. 1979. *Essentials of educational measurement*. (3rd edition). Englewood Cliffs, NJ: Prentice-Hall.
- Ebrahimi, Alireza B. 2018. "Measuring productive collocational knowledge of the most frequent words". *International Journal of Applied Linguistics* 29, 30-43.
- Education group. 2018. Oö. Kinder-Medien-Studie 2018: Das Medienverhalten der 3- bis 10-Jährigen. Land Oberösterreich. <https://www.edugroup.at/innovation/detail/6-ooe-kinder-medien-studie-2018.html>

- Education group. 2019. Oö Jugend-Medien-Studie 2019: Das Medienverhalten der 11- bis 18-Jährigen. Land Oberösterreich. <https://www.edugroup.at/innovation/detail/6-ooe-jugend-medien-studie-2019.html>
- Ehri, Linnea C. 2014. "Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning". *Scientific Studies of Reading* 18, 5–21.
- Ehsanzadeh, Seyed J. 2012. "Depth versus breadth of lexical repertoire: Assessing their roles in EFL students' incidental vocabulary acquisition." *TESL Canada Journal* 29(2), 24-41.
- Ekşi, Gül; Aydın, Hasan. 2013. "What are the students doing 'out' there?: An investigation of out-of-class language learning activities". *AİBÜ Sosyal Bilimler Enstitüsü Dergisi* 13(2), 191-210.
- Elgort, Irina. 2012. "Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test". *Language Testing* 30(2), 253-272.
- Elisson, Dennis. 2017. "Extramural English in the First Grade". MA thesis, Dalarna University.
- Ellis, Nick C.; Simpson-Vlach, Rita; Maynard Carson. 2008. "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL". *TESOL Quarterly* 42(3), 375-396.
- Ellis, Rod. 2009. "Implicit and explicit learning, knowledge and instruction". In Ellis, Rod; Loewen, Shawn; Elder, Catherine; Erlam, Jenefer P.; Reinders, Hayo (eds.). *Implicit and explicit knowledge in second language learning, testing, and teaching*. Bristol: Multilingual Matters, 3-25.
- Enayat, Mostafa J.; Amirian, Seyed M.R. 2020. "The relationship between vocabulary size and depth for Iranian EFL learners at different language proficiency levels". *Iranian Journal of Language Teaching Research* 8(2), 97-114.
- Erman, Britt; Warren, Beatric. 2000. "The idiom principle and the open choice principle". *Text* 20(1), 29 -62.
- Eyckmans, June. 2009. "Toward an assessment of learners' receptive and productive syntagmatic knowledge". In Barfield, Andy; Gyllstad, Henrik (eds.). *Researching collocations in another language: Multiple interpretations*. New York, NY: Palgrave Macmillan, 139-152.
- Fallkvist, Anneli. 2016. "Out-of-school English and the possible effect it has on second language acquisition: A study on how students with different backgrounds acquire the English language outside of school". MA thesis, Stockholm University.
- Fan, May. 2000. "How big is the gap and how to narrow it?: An investigation into the active and passive vocabulary knowledge of L2 learners". *RELJ Journal* 31, 105-119.

- Farrell, Paul. 1990. "Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary". *CLCS Occasional Paper* 25, 1-84.
<https://files.eric.ed.gov/fulltext/ED332551.pdf>
- Faul, Franz; Erdfelder, Edgar; Lang, Albert G.; Buchner, Axel. 2007. "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". *Behavior Research Methods* 39, 175-191.
- Fay, Kristen; Boyd, Michelle. 2010. "Eta-squared". In Kristen, Fay; Boyd Michelle (eds.). *Encyclopedia of Research Design*. SAGE Publications.
<https://dx.doi.org/10.4135/9781412961288>
- Feierabend, Sabine; Plankenhorn, Theresa; Rathgeb, Thomas. 2014. miniKIM: Kleinkinder und Medien. Stuttgart: Medienpädagogischer Forschungsverbund Südwest.
- Ferdous, Tanjila. 2013. "Use of English beyond the classroom wall : A study of undergraduate students' out -of-class English learning activities". BA thesis: BRAC University.
- Field, Andy. 2009. *Discovering Statistics Using IBM SPSS Statistics*. London: SAGE.
- Field, Andy. 2018. *Discovering statistics using IBM SPSS Statistics*. London: SAGE.
- Field, Andy; Miles, Jeremy; Field, Zoe. 2012. *Discovering Statistics Using R*. London: SAGE.
- Firth, John R. 1951. "Modes of meaning". In Palmer, Frank R. (ed.). *Papers in linguistics 1934-1951*. London: Oxford University Press, 190-215.
- Firth, John R. 1957. "A synopsis of linguistic theory". In Palmer, Frank R. (ed.). *Selected papers of J.R. Firth*. London: Oxford University Press, 168-205.
- Firth, John R. 1968 [1952/3]. "Linguistic analysis as a study of meaning". In Palmer, Frank R. (ed.). *Selected papers of J.R. Firth*. London: Oxford University Press, 12-26.
- Fitzpatrick, Tess; Clenton, Jon. 2017. "Making sense of learner performance on tests of productive vocabulary knowledge". *TESOL Quarterly* 51(4), 844-867.
- Fontana, Andrea; Fray, James H. 2005. "The interview". In Denzin, Norman K.; Lincoln, Yvonna S. (eds.). *The Sage Handbook of Qualitative Research* (3rd edition). Thousand Oaks: Sage, 695-728.
- Forsman, Liselott. 2004. "Language, culture and context: Exploring knowledge and attitudes among Finland-Swedish EFL-students with particular focus on extracurricular influence". PhD thesis, Vasa: Åbo Akademi.
- Freeman, Mike. 1999. "The language learning activities of students of ELF and French at two universities". *Language Learning Journal* 19(1), 80-88.

- Freyd, Pamela; Baron Jonathon. 1982. "Individual differences in acquisition of derivational morphology". *Journal of Verbal Learning and Verbal Behavior*, 21, 282-295.
- Gardner, Dee. 2007. "Validating the construct of word in Applied Corpus-based Vocabulary Research: A critical survey". *Applied Linguistics* 28(2), 241-265.
- Gardner, Dee; Davies, Mark. 2014. "A new academic vocabulary list". *Applied Linguistics* 35(3), 305-327.
- Garnier, Melodie; Schmitt, Norbert. 2016. "Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge?". *System* 59, 29-44.
- Gass, Susan M.; Mackey, Alison. 2000. *Stimulated recall methodology in Second Language Research*. Mahwah: Lawrence Erlbaum.
- Ghamarian, Katharina. 2022. "Weaving academic texts: An observation of students' quest for the magic of words in the English language competence programme". In: Berger, Armin; Heaney, Helen; Resnik, Pia; Rieder-Bünemann, Angelika; Savukova, Galina (eds.). *Developing Advanced English Language Competence: A Research-Informed Approach at Tertiary Level*. Springer: English Language Education Band 22, 281-295.
- Gilner, Leah; Morales Frank. 2008. "Corpus-based frequency profiling: Migration to a word list based on the British National Corpus". *The Buckingham Journal of Language and Linguistics* 1, 41-58.
- Goeman, Jelle J.; de Jong, Nivja H. 2018. "How well does the sum score summarize the test? Summability as a measure of internal consistency". *Educational Measurement: Issues and Practice* 37(2), 54-63.
- González-Fernández, Beatriz. 2021. "Vocabulary knowledge: Lexical depth and its relationship with out-of-class exposure". In: Szudarski, Paweł; Barclay, Samuel (eds.). *Vocabulary theory, patterning and teaching*, 21-47.
- González-Fernández, Beatriz; Schmitt, Norbert. 2015. "How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure". *International Journal of Applied Linguistics* 166(1), 94-126.
- González-Fernández, Beatriz; Schmitt, Norbert. 2019. "Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components". *Applied Linguistics* 41(4), 481-505.
- Gorman, Gary E.; Clayton, Peter. 2005. *Qualitative research for the information professional: A practical handbook*. London: Facet.
- Goulden, Robin; Nation, Paul I.S.; Read, John. 1990. "How large can a receptive vocabulary be?". *Applied Linguistics* 11(4), 341-363.

- Graham, Andrew. 2018. *English for academic purposes: A handbook for students*. St. Albans: Critical Publishing.
- Granger, Sylviane. 1998. "Prefabricated patterns in advanced EFL writing: Collocations and formulae". In Cowie, Anothny P. (ed.). *Phraseology: Theory, analysis and applications*. Oxford: OUP, 145–160.
- Grau, Maike. 2009. "Worlds apart? English in German youth cultures and in educational settings". *World Englishes* 28(2), 160–174.
- Green, Rita. 2013. *Statistical analysis for language testers*. New York: Palgrave MacMillan.
- Greenberg, Bradley S.; Sherry, John; Lachlan, Kenneth; Lucas, Kirsten; Holmstrom, Amanda. 2010. "Orientations to video games among gender and age groups". *SAGE* 41(2), 238-259.
- Greidanus, Tine; Nienhus, Lydius. 2001. "Testing the quality of word knowledge in a second language by mean of word associations: Types of distractors and types of associations". *Modern Language Journal* 85(4), 567-577.
- Greidanus, Tine; Bogaards, Paul; van der Linden, Elisatbeth; Nienhuis, Lydius; de Wolf, Tom. 2004. "The construction and validation of a deep word knowledge test for advanced learners of French". In Bogaards, Paul (ed.). *Vocabulary in a second language: selection, acquisition, and testing*. Amsterdam: Benjamins, 191- 227.
- Guttman, Louis. 1944. "A basis for scaling qualitative data". *American Sociological Review* 9, 139-150.
- Gyllstad, Henrik. 2004. "Testing L2 vocabulary: Current test formats in English as a L2 used at Swedish universities". *The department of English in Lund: working papers in linguistics*, 21-40.
- Gyllstad, Henrik. 2007. "Testing English collocations: Developing receptive tests for use with advanced Swedish learners." *Learners Språk- och litteraturcentrum, Lunds universitet*.
- Gyllstad, Henrik. 2009. "Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH". In Barfield, Andy; Gyllstad, Henrik (eds.). *Researching Collocations in Another Language: Multiple Interpretations*. Palgrave Macmillan, 153-170.
- Gyllstad, Henrik. 2013. "Looking at L2 vocabulary knowledge dimensions from an assessment perspective: Challenges and potential solutions". In Bardel, Camilla; Lindqvist, Christina; Laufer, Batia (eds.). *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*. Eurosla Monograph Series 2. European Association of Second Language Acquisition, 11–28. ISBN: 978-1-300-88407-1.

- Gyllstad, Henrik. 2019. "Measuring knowledge of multiword items". In Webb, Stuart (ed.). *Routledge Handbooks in Linguistics, Routledge Handbook of Vocabulary Studies*. Routledge, 387-405.
- Gyllstad, Henrik. 2022. "The Ontogenesis Model : How do multiword units fit in, and are most lexical representations in the L1 really at their optima?". *Bilingualism: Language and Cognition* 25, 216-217.
- Gyllstad, Henrik; McLean, Stuart; Stewart, Jeffrey. 2021. "Using confidence intervals to determine adequate item sample sizes for vocabulary tests : An essential but overlooked practice". *Language Testing* 38, 558-579.
- Gyllstad, Henrik; Schmitt, Norbert. 2018. "Testing formulaic language". In Siyanova-Chanturia, Anna; Pellicer-Sánchez, Ana (eds.). *Second Language acquisition research, understanding formulaic language : A second language acquisition perspective*. Routledge, 174-191.
- Gyllstad, Henrik; Snoder, Per. 2021. "Exploring learner corpus data for language testing and assessment purposes : The case of verb + noun collocations". In Granger, Sylviane (ed.). *Perspectives on the L2 Phrasicon : The view from learner corpora*. Multilingual Matters, 49-71.
- Gyllstad, Henrik; Suhonen, Lari-Valtteri. 2017. "Is attrition a type of learning? : Modelling change in the multilingual mental lexicon". *Linguistic Approaches to Bilingualism* 7, 700-703.
- Gyllstad, Henrik; Vilkaite, Laura; Schmitt, Norbert. 2015. "Assessing vocabulary size through multiple-choice formats : Issues with guessing and sampling rates". *ITL: Instituut Voor Toegepaste Linguïstiek* 166, 278-306.
- Gyllstad, Henrik; Wolter, Brent. 2016. "Collocational processing in light of the phraseological continuum model: Does semantic transparency matter?". *Language Learning* 66(2), 296-323.
- Ha, Althea; Hyland Ken. 2017. "What is technicality?: A technicality analysis model for EAP vocabulary". *Journal of English for Academic Purposes* 28, 35-49.
- Haahr, Mads. 2020. RANDOM.ORG. <http://www.random.org> [last accessed 31.05.2020]
- Hahn, Magdalena. 2017. "Extramural vocabulary acquisition: A survey of students in Viennese vocational business middle schools (HAS)". MA thesis, University of Vienna.
- Halliday, Michael A.K. 1961. "Categories of the theory of grammar". *Word* 17, 241-292.
- Halliday, Michael A.K. 1966. "Lexis as a linguistic level". In Bazell, Charles E.; Catford, John C.; Halliday, Michael A.K.; Robins, Robert H. (eds.). *In memory of J.R. Firth*. London: Longmans, 148-162.

- Hannibal Jensen, Signe. 2017. "Gaming as an English language learning resource among young children in Denmark". *CALICO Journal* 34(1), 1–19.
- Hartley, John. 2020. *Communication, cultural and media studies: The key concepts*. Milton: Routledge.
- Hashimoto, James B. 2021. "Is Frequency enough?: The frequency model in vocabulary size testing". *Language Assessment Quarterly* 18(2), 171-187. DOI: 10.1080/15434303.2020.1860058
- Hatch, Evelyn; Brown Cheryl. 1995. *Vocabulary, Semantics, and Language Education*. New York: Cambridge.
- Hedge, Tricia. 2000. *Teaching and learning in the language classroom*. Oxford: Oxford University Press.
- Helgeson, Joy. 2018. "English in the digital era: Swedish grades 4-6 teachers' use of pupil's extramural English experience of new media". [Unpublished MA thesis]. Dalarna University.
- Henning, Grant. 1987. *A guide to language testing: Development, evaluation and research*. Cambridge: Newbury House.
- Henriksen, Birgit. 1999. "Three dimensions of vocabulary development". *Studies in Second Language Acquisition* 21, 303–317.
- Henry, Alastair; Korp Helena; Sundqvist, Pia; Thorsen Cecilia. 2018. "Motivational strategies and the reframing of English: Activity design and challenges for teachers in contexts of extensive extramural encounters". *TESOL QUARTERLY* 52(2), 247-273.
- Hlebnikovs, Pjotrs. 2018. "Extramural English: Swedish upper secondary students' beliefs on using and learning English outside the classroom". BA thesis, Dalarna University.
- Hong, Ang L.; Hua, Tan K. 2018. "Specificity in English for Academic Purposes (EAP): A corpus analysis of lexical bundles in academic writing". *The Southeast Asian Journal of English Language Studies* 24(2), 82-94.
- Housen, Alex; Pierrard, Michael. 2005. "Investigating instructed Second Language acquisition". In Housen, Alex; Pierrard, Michael (eds.). *Investigations in instructed Second Language Acquisition*. Berlin: Mouton de Gruyter, 1- 30.
- Howarth, Peter. 1996. *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Lexicographica Series Maior, 75. Tübingen: Max Niemeyer.
- Howarth, Peter. 1998a. "Phraseology and second language proficiency". *Applied Linguistics*, 19(1), 24-44.

- Howarth, Peter. 1998b. "The phraseology of learners' academic writing". In Cowie, Anothny P. (ed.). *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press, 161-186.
- Howell, David C. 2013. *Statistical methods for psychology*. (8th edition). Belmont, CA: Wadsworth Cengage Learning.
- Hsu, Wenhua. 2013. "Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list". *Language Teaching Research* 17(4), 454-484.
- Hu, Li-tze; Bentler, Peter M. 1999. "Fit indices in covariance structure modelling: Sensitivity to underparameterized model misspecification". *Psychological Methods* 3(4), 424-453.
- Hughes, Arthur. 2020. *Testing for language teachers*. Cambridge: Cambridge University Press.
- Huibregtse, Ineke; Admiraal, Wilfried; Meara, Paul. 2002. "Scores on a yes-no vocabulary test: correction for guessing and response style". *Language Testing* 19(3), 227-245.
- Hunston, Susanne. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hyland, Fiona. 2004. "Learning autonomously: Contextualising out-of-class English language learning". *Language Awareness* 13(3), 180-202.
- Hyland, Ken. 2016. "General and specific EAP". In: Hyland, Ken; Shaw, Philip (eds.). *The Routledge Handbook of English for Academic Purposes*. London: Routledge, 17- 29.
- Hyland, Ken; Tse, Polly. 2007. "Is there an 'academic vocabulary'?" *TESOL Quarterly* 41(2), 235-253.
- IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp
- Ingvarsdóttir, Hafðís; Jóhannsdóttir, Ásrún. 2018. "Learning and using English: The views of learners at the end of compulsory education". *Educational Linguistics* 34, 79-94.
- Inozu, Julide; Sahinkarakas, Sehnaz; Yumru, Hulya. 2010. "The nature of language learning experiences beyond the classroom and its learning outcomes". *US-China Foreign Language* 8(1), 14-21.
- Irujo, Suzanne. 1993. "Steering clear: Avoidance in the production of idioms". *IRAL* 31(3), 204-219.
- Ishihara, Kenji; Okada, Tae; Matsui Shimpei. 1999. "English vocabulary recognition and production: A preliminary survey report". *Doshisha Studies in Language and Culture*, 2(1), 143-175.

- Jakonen, Teppo. 2014. "Building bridges: How secondary school pupils bring their informal learning experiences into a content and language integrated (CLIL) classroom". *APPLES – Journal of Applied Language Studies* 8(1), 7–28.
- Jeon, Eun H. 2015. "Multiple Regression". In Plonsky, Luke (ed.). *Advancing quantitative methods in second language research*. New York: Routledge, 131-158.
- Jiang Nan; Nekrasova, Tatiana M. 2007. "The processing of formulaic sequences by second language speakers". *Modern Language Journal* 91(3), 433-445.
- Jiménez Catalán, Rosa; Ruiz de Zarobe, Yolanda; Cenoz, Jason. 2006. "Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language". *VIEWS (Vienna English Working Papers)* 15(3), 23-27.
- Jóhannsdóttir, Ásrún. 2017. "English exposure and vocabulary proficiency at the onset of English instruction". In Arnbjörnsdóttir, Birna; Ingvarsdóttir, Hafdís (eds.). *Language development across the life span: The impact of English on education and work in Iceland*. Cham: Springer, 57–78.
- Kalaja, Paula; Alanen, Riikka; Palviainen, Åsa; Dufva, Hannele. 2011. "From milk cartons to English roommates: Content and agency in L2 learning beyond the classroom". In Benson, Phil; Reinders, Hayo (eds.). *Beyond the language classroom*. Houndsmill, Basingstoke: Palgrave Macmillan, 47–58.
- Kamarudin, Rafidah; Abdullah, Shazila; Abdul Aziz, Roslina. 2020. "Examining ESL learners' knowledge of collocations". *International Journal of Applied Linguistics & English Literature* 9(1), 1-6.
- Karlsson, Monica. 2013. "Quantitative and qualitative aspects of advanced learners' L1 and L2 mastery of polysemous words". *Journal of Language and Communication in Business* 51, 79-112.
- Kaur, Naginder. 2015. "Making meaning of vocabulary learning: Seizing opportunities at opportune moments". *GEMA Online® Journal of Language Studies* 15(2), 1–16.
- Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London: Longman.
- Khani, Reza; Tazik, Khalil. 2013. "Towards the development of an academic word list for Applied Linguistics research articles". *RELJ Journal* 44(2), 209-232.
- Kit, Chunyu; Liu, Xiaoyue. 2008. "Measuring mono-word termhood by rank difference via corpus comparison". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 14(2), 204-229.
- Kline, Paul. 1999. *The handbook of psychological testing* (2nd edition). London: Routledge.
- Knight, Tracey. 2007. "Beyond the classroom walls: A study of out of class English use by adult community college ESL students". MA thesis, Portland State University.

- Knoch Ute. 2014. "Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organisation (ICAO) rating scale". *English for Specific Purposes* 33, 77-86.
- Koivistoinen, Hilka A. 2015. "Crossing geographies of language learning: The case of 'a successful pupil'". *Classroom Discourse* 6(1), 20–32.
<https://doi.org/10.1080/19463014.2014.893895>
- Kremmel, Benjamin. 2016. "Word families and frequency bands in vocabulary tests: Challenging conventions". *TESOL QUARTERLY* 50(4), 976- 987.
- Kremmel, Benjamin. 2020. "Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey". *Language Assessment Quarterly* 17(1), 100-120.
- Kremmel, Benjamin; Harding, Luke. 2020."Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey". *Language assessment quarterly* 17(1), 100-120.
- Kremmel, Benjamin; Schmitt, Norbert. 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.
- Kuppens, An H. 2010. "Incidental foreign language acquisition from media exposure". *Learning, Media and Technology* 35(1), 65–85.
- Kusyk, Meryl. 2017. "The development of complexity, accuracy and fluency in L2 written production through informal participation in online activities". *Calico Journal* 34, 75-96.
- Kuure, Leena. 2011. "Places for learning: Technology-mediated language learning practices beyond the classroom". In Benson, Phil; Reinders, Hayo (eds.). *Beyond the language classroom*. Houndsmill, Basingstoke: Palgrave Macmillan, 35–46.
- Kwary, Deny A. 2011. "A hybrid method for determining technical vocabulary". *System* 39, 175-185.
- Lado, Robert. 1961. *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longmans.
- Lai, Chun. 2016. "Perceiving and traversing in-class and out-of-class learning: accounts from foreign language learners in Hong Kong". *Innovation in Language Learning and Teaching* 9(3), 265-284.
- Lai, Chun; Gu, Mingyue. 2011. "Self-regulated out-of-class language learning with technology". *Computer Assisted Language Learning* 24(4), 317–335.

- Lai, Chun; Hu, Xiao; Lyu, Boning. 2018. "Understanding the nature of learners' out-of-class language learning experience with technology". *Computer Assisted Language Learning* 31(1-2), 114–143.
- Lai, Chun; Yeung, Yuk; Hu, Jingjing. 2016. "University student and teacher perceptions of teacher roles in promoting autonomous language learning with technology outside the classroom". *Computer Assisted Language Learning* 29(4), 703–723.
- Lai, Chun; Zheng, Dongping. 2018. "Self-directed use of mobile devices for language learning beyond the classroom". *ReCALL* 30(3), 299–318.
- Lai, Chun; Zhu, Weimin; Gong, Gang. 2015. "Understanding the quality of out-of-class English learning". *TESOL Quarterly* 49(2), 278–308.
- Lamb, Martin. 2002. "Explaining successful language learning in difficult circumstances". *Prospect* 17(2), 35–52.
- Lamb, Martin. 2004a. "Integrative motivation in a globalizing world". *System* 32(1), 3–19.
- Lamb, Martin. 2004b. "'It depends on the students themselves': Independent language learning at an Indonesian state school". *Language, Culture and Curriculum* 17(3), 229–245.
- Lancaster, Nina K. 2018. "Extramural exposure and language attainment: The examination of input-related variables in CLIL programmes". *Porta Linguarum* 29, 91–114.
- Larson-Hall, Jenifer. 2016. *A guide to doing statistics in second language research using SPSS and R*. (2nd edition). New York: Routledge.
- Laufer, Batia. 1990. "Why are some words more difficult than others?: Some intralexical factors that affect the learning of words". *International Review of Applied Linguistics* 28, 293–307.
- Laufer, Batia. 1998. "The development of passive and active vocabulary in a second language: Same or different?" *Applied Linguistics* 19(2), 255–271.
- Laufer, Batia. 2005. "Lexical frequency profiles: From Monte Carlo to the real world". *Applied Linguistics* 26(4), 582–588.
- Laufer, Batia; Elder, Cathie; Hill, Kathryn; Congdon, Peter. 2004. "Size and strength: do we need both to measure vocabulary knowledge?". *Language Testing* 21(2), 202–226.
- Laufer, Batia; Goldstein, Zahava. 2004. "Testing vocabulary knowledge: Size, strength, and computer adaptiveness". *Language Learning* 54(3), 399–436.
- Laufer, Batia; Hulstijn, Jan H. 2001. "Incidental vocabulary acquisition in a second language: The construct of task-induced involvement". *Applied Linguistics* 22(1), 1–26.

- Laufer, Batia; Nation, Paul I.S. 1995. "Vocabulary size and use: Lexical richness in L2 written production". *Applied Linguistics* 16(3), 307-322.
- Laufer, Batia; Nation, Paul I.S. 1999. "A vocabulary-size test of controlled productive ability". *Language Testing* 16(1), 33-51.
- Laufer, Batia; Nation Paul I.S. 2001. "Passive vocabulary size and speed of meaning recognition: Are they related?". *EUROSLA Yearbook* 1, 7-28.
- Laufer, Batia; Ravenhorst-Kalovski, Geke C. 2010. "Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension". *Reading in a Foreign Language* 22, 15-30.
- Laufer, Batia; Waldman, Tina. 2011. "Verb-Noun collocations in second language writing: A corpus analysis of learners' English". *Language Learning* 61(2), 647- 672.
- Lawrence, Joshua F.; Hagen, Aste M.; Hwang, Jin K.; Lin, Grace; Lervag, Arne. 2019. "Academic vocabulary and reading comprehension: Exploring the relationships across measures of vocabulary knowledge". *Read Writ* 32, 285-306.
- Lee, Ju S. 2019a. "Informal digital learning of English and second language vocabulary outcomes: Can quantity conquer quality?". *British Journal of Educational Technology* 50(2), 767-778.
- Lee, Ju S. 2019b. "Quantity and diversity of informal digital learning of English". *Language Learning & Technology* 23(1), 114-126.
- Lee, Ju S.; Dressman, Mark. 2018. "When IDLE hands make an English workshop: Informal digital learning of English and language proficiency". *TESOL Quarterly* 52(2), 435-445.
- Lee Ju S.; Sylvén Liss K. 2021. "The role of informal digital learning of English in Korean and Swedish EFL learners' communication behaviour". *British Journal of Educational Technology* 52, 1279-1296.
- Lee Ju S.; Taylor Tim. 2022." Positive psychology constructs and Extramural English as predictors of primary school students' willingness to communicate." *Journal of Multilingual and Multicultural Development*, 1-19. DOI: 10.1080/01434632.2022.2079650
- Lee, Sunjung. 2020. "Examining the roles of aptitude, motivation, strategy use, language processing experience, and gender in the development of the breadth and depth of EFL learners' vocabulary knowledge". *SAGE Open* 10(4), 1-20.
- Lefever, Samúel. 2010. "English skills of young learners in Iceland: "I started talking English when I was 4 years old. It just bang... just fall into me"". *Ráðstefnurit Netlu – Menntakvika*, 1-17.

- Lei, Lei; Liu, Dilin. 2016. "A new medical academic word list: A corpus-based study with enhanced methodology". *Journal of English for Academic Purposes* 22, 42-53.
- Lehtonen, Tuula. 2011. "Out-of-classroom language learning". Pitkänen, Kari K., Jokinen, Jaana, Karjalainen, Sinikka, Karlsson, Leena, Lehtonen, Tuula, Matilainen, Mirjami, Niedling, Christian; Siddall, Roy (eds.). Helsinki: Helsingin yliopiston kielikeskus, 69-84.
- Li, Jie; Schmitt, Norbert. 2009. "The acquisition of lexical phrases in academic writing: A longitudinal case study". *Journal of Second Language Writing* 18, 85-102.
- Li, Miao; Kirby, John R. 2015. "The effects of vocabulary breadth and depth on English reading." *Applied Linguistics* 35(6), 611-634.
- Li, Yan; Hafner, Christoph A. 2021. "Mobile-assisted vocabulary learning: Investigating receptive and productive vocabulary knowledge of Chinese EFL learners." *ReCALL*, 1-15. doi:10.1017/S0958344021000161
- Liao, Yan D.; Fukuya, Yoshinori J. 2004. "Avoidance of phrasal verbs: The case of Chinese learners of English". *Language Learning* 54(2), 193-226.
- Lindgren, Eva; Muñoz, Carmen. 2013. "The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension". *International Journal of Multilingualism* 10(1), 105-129.
- Lindsay, Shane; Gaskell, Gareth M. 2010. "A complementary systems account of word learning in L1 and L2". *Language Learning* 60, 45-63.
- Littlewood, William; Liu, Ngan F. 1996. *Hong Kong students and their English*. Hong Kong: Hong Kong University.
- Liu, Dilin. 2012. "The most frequently-used multi-word constructions in academic written English: A multi-corpus study". *English for Specific Purposes* 31, 25-35.
- Liu, Dilin; Lei, Lei. 2020. "Technical vocabulary." In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 111-124.
- Liu, Na; Lin, Chih-Kai; Wiley, Terrence G. 2016. "Learner views on English and English language teaching in China". *International Multilingual Research Journal* 10(2), 137-167.
- Liu, Xianghu. 2014. "Students' perceptions of autonomous out-of-class learning through the use of computers". *English Language Teaching* 7(4), 74-82.
- Llach, Pilar A.; Gallego, Melania T. 2009. "Examining the relationship between receptive vocabulary size and written skills of primary school learners". *Journal of the Spanish Association of Anglo-American Studies* 31(1), 129-147.

- Loewen, Shawn; Gonulal, Talip. 2015. "Exploratory factor analysis and principal components analysis". In Plonsky, Luke (ed.). *Advancing quantitative methods in second language research*. New York: Routledge, 182-212.
- Logan, Kenneth J. 2018. "Exploring the relationship between polysemous word knowledge, executive function, and reading comprehension among early adolescents". PhD thesis, New York University.
- Logan, Kenneth J.; Kieffer, Michael J. 2017. "Evaluating the role of polysemous word knowledge in reading comprehension among bilingual adolescents." *Reading and Writing* 30(8), 1687-1704.
- Logan, Kenneth J.; Kieffer, Michael J. 2021. "Investigating the longitudinal and concurrent relationships between polysemous word knowledge and reading comprehension among Spanish-English bilingual middle school students". *Reading and Writing* 34, 301-335.
- Lundahl, Bo. 2009. *Engelsk språkdidaktik: Texter, kommunikation, språkutveckling*. Lund: Studentlitteratur.
- Luxton, Julie; Fry, Juliet; Coxhead, Averil. 2017. "Exploring the knowledge and development of academic English vocabulary of students in New Zealand secondary schools". *Teaching and Learning* 1, 12-22.
- Luzon, María-José. 2018. "Variation in academic writing practices: The case of reporting verbs in English medium research articles". *Iberic a* 36, 171-194.
- Lyriqkou, Christina. 2019. "Not to be overlooked: agency in informal language contact". *Innovation in Language Learning and Teaching* 13(3), 237-252.
- Ma, Yu-han; Lin, Wen-ying. 2015. "A study on the relationship between English reading comprehension and English vocabulary knowledge." *Education Research International*, 1-14.
- Maby, Mark. 2016. "An investigation of L2 English learners' knowledge of polysemous word senses". PhD thesis, Cardiff University.
- Mahony, Diana; Singson, Maria; Mann Virginia. 2000. "Reading ability and sensitivity to morphological relations". *Reading and Writing: An Interdisciplinary Journal* 12, 191-218.
- Mair Patrick; Wilcox Rand; Patil, Indrajeet. 2005. *WRS2: Wilcox robust estimation and testing*. R package version (Version 0.4-0). Retrieved from <http://cran.r-project.org/package=WRS2>

- Makhoul, Baha. 2017. "Investigating Arabic academic vocabulary knowledge among middle school pupils: Receptive versus productive knowledge." *J Psycholinguist Research* 46, 1053-1065.
- Manfred, Wu M. 2012. "Beliefs and Out-of-class language learning of Chinese-speaking ESL learners in Hong Kong". *New Horizons in Education* 60(1), 35-52.
- Marchman, Virginia A.; Fernald, Anne; Hurtado, Nereyda. 2010. "How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish–English bilinguals." *J. Child Lang* 37, 817-840.
- Martínez Agudo, Juan de Dios. 2019. "Which instructional programme (EFL or CLIL) results in better oral communicative competence?: Updated empirical evidence from a monolingual context". *Linguistics and Education* 51, 69-78.
- Martinez, Ron. 2011. "Putting a test of multiword expressions to a test". Paper presented at the *IATEFL Testing, Evaluation and Assessment SIG*. University of Innsbruck: September 16. Retrieved from <https://ufpr.academia.edu/RonMartinez/Talks>
- Martinez, Ron; Schmitt, Norbert. 2012. "A phrasal expressions list". *Applied Linguistics* 33(3), 299-320.
- Masrai, Ahmed; Milton, James. 2018. "Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement." *Journal of English for Academic Purposes* 31, 44-57.
- Mayring, Phillip. 2007. *Mixed Methodology in psychological research*. Boston: Brill.
- McBride-Chang, Catherine; Wagener, Richard K.; Muse Andrea; Chow, Bonnie W.Y.; Shu, Hua. 2005. "The role of morphological awareness in children's vocabulary acquisition in English". *Applied Psycholinguistics* 26, 415-435.
- McCracken, Grant. 1988. *The long interview*. Newbury Park: Sage.
- McLeod, Saul A. 2019. "What does effect size tell you?". *Simply Psychology*. Retrieved from: <https://www.simplypsychology.org/effect-size.html>.
- McKeown, Margaret G.; Crosson Amy C.; Moore, Debra W.; Beck Isabell L. 2018. "Word knowledge and comprehension effects of an academic vocabulary intervention for middle school students". *American Educational Research Journal* 55(3), 572-616.
- Meara, Paul M. 1999. "Lexis: Acquisition". In Spolsky, Bernard (ed.). *Concise encyclopedia of educational linguistics*. New York: Elsevier, 565–567.
- Meara, Paul M.; Milton, James. 2003. *X_Lex, The Swansea Levels Test*. Newbury: Express.
- Meara, Paul M. 2009. *Connected words : word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins Publishing Co.

- Meara, Paul M. 2005. "Designing vocabulary tests for English, Spanish and other languages". In Butler, Christopher, Gómez-González, María & Doval Suárez, Susana M. (eds.), *The dynamics of language use*. Amsterdam: Benjamins, 271-286.
- Meara, Paul M. 2010. *EFL Vocabulary Tests*. Washington: ERIC.
- Meara, Paul. 2015. "V_YesNo: a Yes/No vocabulary test for English (v1.01)". *Swansea: lognostics*. Retrieved from http://www.lognostics.co.uk/tools/V_YesNo/V_YesNo.htm
- Meara, Paul M.; Buxton, Barbara. 1987. "An alternative to multiple choice vocabulary tests". *Language Testing* 4(2), 142-154.
- Meara, Paul M.; Fitzpatrick, Tess. 2000. "Lex30: an improved method of assessing productive vocabulary in an L2". *System* 28, 19-30.
- Meara, Paul M. and Jones, Glyn. 1990. *Eurocentres Vocabulary Size Tests*. Zurich: Eurocentres Learning Service.
- Meara, Paul M.; Miralpeix, Imma. 2017. *Tools for researching vocabulary*. Bristol: Multilingual Matters.
- Meara, Paul M.; Wolter, Brent. 2004. "V_Links: Beyond vocabulary depth". *Angles on the English speaking world* 4, 85-110.
- Medienpädagogischer Forschungsverbund Südwest. 2019. *JIM Studie 2019: Jugend, Information, Medien: Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger*. Retrieved from: https://www.mpfs.de/fileadmin/files/Studien/JIM/2019/JIM_2019.pdf
- Melčuk, Igor A. 1998. "Collocations and lexical functions". In Cowie, Anthony P. (ed.). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press, 24-53.
- Melka, Francine. 2011. "Receptive vs. productive aspects of vocabulary". In Schmitt, Norbert; McCarthy, Michael (eds.). *Vocabulary: Description, acquisition, and pedagogy*. New York: Cambridge University Press, 84-102.
- Messick, Samuel. 1989. "Validity". In Linn, Richard L. (ed.). *Educational Measurement*. Phoenix: The Oryx Press, 13-104.
- Michalec, David; Henninger, Nathan. 2011. "Expressive One-Word Picture Vocabulary Test". In: Kreutzer, Jeffrey S., DeLuca, John; Caplan, Bruce (eds.). *Encyclopedia of Clinical Neuropsychology*. New York: Springer https://doi.org/10.1007/978-0-387-79948-3_1544
- Miller, Megan; Hegelheimer, Volker. 2006. "The SIMs meet ESL incorporating authentic computer simulation games into the language classroom". *Interactive Technology & Smart Education* 4, 311-328.

- Milton, James. 2009. "Measuring Second Language vocabulary acquisition." *Studies in Second Language Vocabulary Acquisition* 33(3), 471-472.
- Milton, James; Fitzpatrick, Tess. 2014. *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillian.
- Miralpeix, Imma. 2007. "Testing receptive vocabulary size: X_Lex and Y_Lex". Paper presented at the *29th Language Testing Research Colloquium*, Barcelona.
- Miralpeix, Imma. 2020. "L1 and L2 vocabulary size and growth". In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 189-206.
- Miralpeix, Imma; Muñoz, Carmen. 2018. "Receptive vocabulary size and its relationship to EFL language skills". *IRAL* 56(1), 1-24.
- Mirmán Flores, Ana; García Jiménez, Eduardo. 2018. "The influence of family environment on exposure to English among Spanish secondary school students". *Estudios sobre Educación* 34, 283–306.
- Miščin, Evelina. 2015. "Collocational competence of primary and secondary school students." *ExELL* 3(1), 8-25.
- Molenaar, Ivo W.; Sijtsma, Klaas. 2000. *MPS5 for Windows. A Program for Mokken Scale Analysis of polytomous items*. iecProGAMMA.
- Moncrief, Robert. 2011. "Out-of-Classroom Language Learning". In Pitkänen, Kari, Jokinen, Jaana, Karjalainen, Sinikka, Karlsson, Leena, Lehtonen, Tuula, Matilainen, Mirjami, Niedling, Christian; Siddall, Roy (eds.). *Language Centre Publications*. Helsinki: Helsingin yliopiston kielikeskus, 107-118 22.
- Mondria, Jan-Arjen; Wiersma, Boukje. 2018. "Receptive, productive, and receptive + productive L2 vocabulary learning: what difference does it make?". In: Bogaards, Paul; Laufer, Batia (eds.). *Vocabulary in second language*. John Benjamins, 79-100.
- Montero Perez, Maribel; Peters, Elke; Clarebout, Geraldine; Desmet, Piet. 2014. "Effects of captioning on video comprehension and incidental vocabulary learning". *Language Learning and Technology* 18(1), 118–141. 10125/44357
- Moon, Rosamund. 2011. "Vocabulary connections: Multi-word items in English". In Schmitt, Norbert; McCarthy, Michael (eds.). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 40-63.
- Morimoto, Shun; Loewen, Shawn. 2007. "A comparison of the effects of image-schema-based instruction and translation-based instruction on the acquisition of L2 polysemous word". *Language Teaching Research* 11(3), 347-372.

- Mudraya, Olga. 2006. "Engineering English: A lexical frequency instructional model". *English for Specific Purposes* 25(2), 235-256.
- Mukundan, Jayakaran; Khojasteh, Laleh; Pearson, Nigel. 2009. "Exploring the language learning materials used out-of-class by Malaysian TESL students and students of TBFL". *Indonesian Journal of English Language Teaching* 5(1), 40–56.
- Muñoz, Carmen. 2012. "The significance of intensive exposure as a turning point in learners' histories". In Muñoz, Carmen (ed.). *Intensive exposure experiences in second language learning*. Bristol: Multilingual Matters, 141–160.
- Nachtigall, Christoff; Kroehne, Ulf; Funke, Friedrich; Steyer, Rolf. 2003. "(Why) should we use SEM?: Pros and cons of structural equation modelling". *Methods of Psychological Research Online* 8(2), 1-22.
- Nagy, William. 1997. "On the role of context in first- and second-language vocabulary learning". In Schmitt, Norbert; McCarthy Michael (eds.). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 64-83.
- Nakata, Tatsuya. 2008. "English Vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning". *ReCALL* 20(1), 3-20.
- Nassaji, Hossein. 2006. "The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success". *Modern Language Journal* 90(3), 387-401.
- Nation, Paul I.S. 1983. "Testing and teaching vocabulary". *Guidelines* 5, 12-25.
- Nation, Paul I.S. 1990. *Teaching and Learning Vocabulary*. New York: Newbury.
- Nation, Paul I.S. 1995. "Testing and Teaching Vocabulary". *Guidelines* 5(1), 12-25.
- Nation, Paul I.S. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, Paul I.S. 2004. "A study of the most frequent word families in the British National Corpus". In Bogaards, Paul; Laufer, Batia (eds.). *Vocabulary in a second language*. Amsterdam: John Benjamins, 3-14.
- Nation, Paul I.S. 2006. "How large a vocabulary is needed for reading and listening?". *Canadian Modern Language Review* 63(1), 59-82.
- Nation, Paul I.S. 2012. *The BNC/COCA word family lists*. Retrieved from <http://www.victoria.ac.nz/lals/about/staff/paul-nation>
- Nation, Paul I.S. 2013. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

- Nation, Paul I.S. 2020. "The different aspects of word knowledge". In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 15-29.
- Nation, Paul I.S. 2021. "Critical commentary : Thoughts on word families". *Studies in Second Language Acquisition* 43, 969-972.
- Nation, Paul I.S.; Beglar, David. 2007. "A vocabulary size test". *The Language Teacher* 31(7), 9-12.
- Nation, Paul I.S.; Coxhead, Averil. 2012. "Special purposes vocabulary". In Chapelle, Carol A. (ed.). *Encyclopedia of applied linguistics*. Hoboken: Wiley Blackwell, 252-267.
- Nation, Paul I.S.; Hwang, Kyongho. 1995. "Where would general service vocabulary stop and special purposes vocabulary begin?". *System* 23(1), 35-41.
- Nation, Paul I.S.; Waring, Robert. 2011. "Vocabulary size, text coverage and word lists". In: Schmitt, Norbert; McCarthy, Michael (eds.). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 6-19.
- Nation, Paul I.S.; Webb, Stuart. 2011. *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage.
- Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.
- Nesselhauf, Nadja. 2009. "Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties". *English World-Wide* 30, 1-26.
- Nguyen, Thi My Hang; Webb, Stuart. 2017. "Examining second language receptive knowledge of collocation and factors that affect learning". *Language Teaching Research* 21(3), 298-320.
- Nightingale, Richard. 2016. "The effect of out-of-school media contact on language attitudes in multilingual adolescents: A complex psycho-sociolinguistic system". PhD thesis, Castelló de la Plana: Jaume I University.
- Nihayra, Leona L.; van Koert, Margreet J.; van der Molen, Maurits W.; Rispen, Judith E.; Tijms, Jurgen; Snellings, Patrick. 2021. "Explaining individual differences in young English language learners' vocabulary knowledge: The role of Extramural English exposure and motivation". *System* 96, 1-13.
- Nilsson, Anna. 2013. "Extramural English or school?: A quantitative study of what factors influence Swedish Senior High School Students' variety of English". BA thesis, Upsala Universitet.
- Nirattisai, Supika; Palanukulwong, Thanyapa. "A new measure for Thai students' vocabulary size: The English-Thai version of vocabulary size test". *Journal of English studies* 14(14), 155-181.

- Nizonkiza, Déogratias. 2015. "Measuring receptive collocational competence across proficiency levels". *Stellenbosch Papers in Linguistics* 44, 125- 146.
- Nizonkiza, Déogratias. 2016. "First-year university students' receptive and productive use of academic vocabulary". *Stellenbosch Papers in Linguistics* 45, 169-187.
- Nizonkiza, Déogratias; van den Berg, Karien. 2014. "The dimensional approach to vocabulary testing: What can we learn from past and present practices?". *Stellenbosch Papers in Linguistics Plus* 43, 45-61. doi: 10.5774/43-0-169
- Noori, Sadegh; Gholami, Hamid; Rajabi, Seifodin. 2014. "The development of the receptive and productive knowledge of collocations through incidental learning vs. intentional learning". *Modern journal of language teaching methods* 4(2), 146-153.
- Nugrahenny, Zacharias T. 2012. *Qualitative research methods for second language education: a coursebook*. Newcastle upon Tyne: Cambridge Scholars Pub
- Nurweni, Ari; Read, John. 1999. "The English vocabulary knowledge of Indonesian university students". *English for specific purposes* 18(2), 161-175.
- Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Olsson, Eva. 2011. "Everything I read on the internet is in English: On the impact of extramural English on Swedish 16-year-old pupils' writing proficiency". *ROSA* 15. Göteborg. Retrieved from: http://www.kultur.gu.se/digitalAssets/1325/1325748_olsson-eva-lic.pdf
- Olsson, Eva. 2016. "On the impact of extramural English and CLIL on productive vocabulary". PhD thesis, Göteborg: University of Gothenburg.
- Olsson, Eva; Sylvén, Liss K. 2015. "Extramural English and academic vocabulary: A longitudinal study of CLIL and non-CLIL students in Sweden". *Apples – Journal of Applied Language Studies* 9(2), 77–103.
- Oppenheim, Angela N. 1992. *Questionnaire design, interviewing and attitude measurement: New edition*. London: Pinter.
- Ordóñez, Claudia L.; Carlo, Maria S.; Snow, Catherine E.; McLaughlin, Barry. 2002. "Depth and breadth of vocabulary in two languages: Which vocabulary skills transfer?". *Journal of Educational Psychology* 94(4), 719-728.
- Orhon, Yelda. 2018. "An investigation of out-of-class language activities of tertiary-level EFL learners". *Education Reform Journal* 3(1), 1-14.
- Orosz, Andrea. 2009. "The growth of young learners' English vocabulary size". In: Nikolov, Marianne (ed.). *Early Learning of Modern Foreign Languages*. Channel View Publications Ltd, 181-194.

- Oxford University Press. (n.d.) In The Oxford Learners Dictionaries. Retrieved April 2, 2020 from <https://www.oxfordlearnersdictionaries.com/>
- Oxford University Press. (n.d.). In Oxford English Dictionary. Retrieved April 3, 2020 from <https://www.oed.com/>
- Özdem Ertürk, Zeynep. 2017. "The effects of receptive and productive learning tasks on EFL learners' knowledge of collocation and meaning". *Eurasian Journal of Applied Linguistics* 3(2), 59-73.
- Ozturk, Meral. 2015. "Vocabulary growth of the advanced EFL learner". *The Language Learning Journal* 43(1), 94-109.
- Ozturk, Meral. 2016. "Second language vocabulary growth at advanced level". *The Language Learning Journal* 44(1), 6-16.
- Palmer, Harold. 1931. *First interim report on vocabulary selection*. Tokyo: Kaitakusha.
- Palmer, Harold. 1933. *Second interim report on English collocations*. Tokyo: Kaitakusha
- Paltridge, Brian; Starfield, Sue; Coxhead, Averil. 2017. "Multi-word units and metaphor in ESP". In Paltridge, Brian; Starfield, Sue; Coxhead, Averil. *Vocabulary and English for Specific Purposes research*. London: Routledge, 45-62.
- Paquot, Magali; Granger, Sylviane. 2012. "Formulaic language in learner corpora". *Annual Review of Applied Linguistics* 32, 130-149.
- Paquot, Magali. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum International Publishing Group.
- Paribakht, Sima, T.; Wesche, Marjorie. 1993. "Reading comprehension and second language development in a comprehension-based EFL program". *TESL Canada Journal* 11(1), 9-29.
- Patton, Michael Q. 2002. *Qualitative research and evaluation methods*. (3rd edition). Thousand Oaks: Sage.
- Pearson, Nigel. 2004. "The idiosyncrasies of out-of-class language learning: A study of mainland Chinese students studying English at tertiary level in New Zealand". *Proceedings of the Independent Learning Conference 2003: Independent Learning Association*. Retrieved from: <http://www.independentlearning.org/proceedings-melbourne-2003.html>.
- Perfetti, Charles A. 2007. "Reading ability: Lexical quality to comprehension". *Scientific Studies of Reading* 11, 357-383.
- Perfetti, Charles A.; Hart, Lesley. 2001. "The lexical basis of comprehension skill". In Gorfein, David S. (ed.). *On the consequences of meaning selection: Perspectives on resolving*

lexical ambiguity. American Psychological Association, 67–86.
<https://doi.org/10.1037/10459-004>

- Perfetti, Charles A.; Hart, Lesley. 2002. "The lexical quality hypothesis". In Verhoeven, Ludo; Elbro, Carsten; Reitsma, Pieter (eds.). *Precursors of Functional Literacy-Studies in Written Language and Literacy* 11. Amsterdam/Philadelphia: John Benjamins Publishing Company, 67-86.
- Perinán-Pascual, Carlos. 2015. "The underpinnings of a composite measure for automatic term extraction: The case of SRC". *Terminology* 21, 151-179.
- Persson, Liv; Prins, Tineke. 2012. "Learning English inside and outside the classroom". In De Jong, Nel; Juffermans, Kasper; Keijzer, Merel; Rasier, Laurent (eds.). *Papers of the Anéla 2012 applied linguistics conference*. Delft: Uitgeverij Eburon, 3-13.
- Peruzzo Katia. 2014. "Term extraction and management based on event templates: An empirical study on an EU corpus". *Terminology* 20, 151-170.
- Peters, Elke. 2014. "The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations". *Language Teaching Research* 18(1), 75-94.
- Peters, Elke. 2018. "The effect of out-of-class exposure to English language media on learners' vocabulary knowledge". *ITL - International Journal of Applied Linguistics* 169(1), 142–168.
- Peters, Elke. 2019. "The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input". *TESOL Quarterly* 53(4), 1008-1032.
- Peters, Elke; Heynen, Eva; Puimège, Eva. 2016. "Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions". *System* 63, 134-148.
- Peters, Elke; Noreillie, Ann-Sophie; Heylen, Kris; Bulté, Bram; Desmet, Piet. 2019. "The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages". *Language Learning* 69(3), 747-782.
- Peters, Elke; Velghe, Tom; Van Rompaey, Tinne. 2015. "A post-entry English and French vocabulary size for Flemish learners". *Paper presented at EALTA*. Copenhagen, Denmark.
- Peters, Elke; Velghe, Tom; Van Rompaey, Tinne. 2019. "The VocabLab Test: The development of an English and French vocabulary test". *International Journal of Applied Linguistics* 170(1), 53-78.
- Pham, Giang; Khonert, Kathryn. 2014. "A longitudinal study of lexical development in children learning Vietnamese and English". *Child Development* 85(2), 767-782.

- Pickard, Nigel. 1996. "Out-of-class language learning strategies". *ELT Journal* 50(2), 150-159.
- Pill, Thomas J. H. 2001. "Adult learners' perceptions of out-of-class access to English". MA thesis, University of Hong Kong.
- Pellicer-Sánchez, Ana; Schmitt, Norbert. 2010. "Incidental vocabulary acquisition from an authentic novel: Do things fall apart?". *Reading in a Foreign Language* 22, 31-55.
- Plag, Ingo. 2003. *Word-Formation in English*. Cambridge: Cambridge University Press.
- Plonsky, Luke. 2015. "Statistical power, p values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research". In Plonsky, Luke (ed.). *Advancing quantitative methods in second language research*. New York: Routledge, 23-45.
- Popham, James W. 2000. *Modern Education Measurement* (3rd edition). Boston: Allyn & Bacon.
- Porte, Graeme K. 2010. *Appraising research in second language learning: A practical approach to critical analysis of quantitative research* (2nd edition). Amsterdam: John Benjamins.
- Puimège, Eva; Peters, Elke. 2019. "Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related factors". *Language Learning* 69(4), 943-977.
- Qian, David. 2002. "Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective". *Language Learning*, 52(3), 513-536.
- Qian, David; Schedl, Mary. 2004. "Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance". *Language Testing* 21(1), 28-52.
- Rafidah Kamarudin, Rafidah; Abdullah; Abdul Aziz, Roslina. 2020. "Examining ESL learners' knowledge of collocations". *International Journal of Applied Linguistics & English Literature* 9(1), 1-6.
- Ranalli, Jim. 2008. "Learning English with The Sims: exploiting authentic computer simulation games for L2 learning". *Computer Assisted Language Learning* 21(5), 441–455.
- Rasztovits, Dominik. 2020. "Extramural English and motivation: Comparing and contrasting language learners' motivation for video games and films". PhD thesis, University of Vienna.
- Read, John. 1993. "The development of a new measure of L2 vocabulary knowledge". *Language Testing* 10, 355-371.

- Read, John. 1995. "Validating the Word Associates Format as a measure of depth of vocabulary knowledge". Paper presented at the 17th Language Testing Research Colloquium, Long Beach, CA.
- Read, John. 1998. "Validating a test to measure depth of vocabulary knowledge". In Kunnan, Antony (ed.). *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 41–60.
- Read, John. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Read, John. 2004. *Word associates test (Version 4.0.)*. Retrieved from: <http://www.lexutor.ca/>
- Read, John. 2010. *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Reinders, Hayo; Benson, Phil. 2017. "Thinking Allowed: Research agenda: Language learning beyond the classroom". *Language Teaching*, 50(4), 561–578.
- Reinders, Hayo; Wattana, Sorada. 2012. "Talk to Me! Games and Students' Willingness to Communicate". In Reinder, Hayo (ed.). *Digital Games in Language Learning and Teaching*. Basingstoke : Palgrave Macmillan, 156-188.
- Revier, Robert L. 2009. "Evaluating a new test of whole English collocations". In Barfield, Andy; Gyllstad, Henrik (eds.). *Researching collocations in another language: Multiple interpretations*. New York, NY: Palgrave Macmillan, 125–138.
- Revier, Robert L. 2014. "Testing knowledge of whole English collocations available for use in written production: Developing tests for use with intermediate and advanced Danish learners". Unpublished PhD thesis, Aarhus University.
- Richard, Gail J.; Hanner, Mary A. 1995. *Language Processing Test–Revised*. East Moline, IL: LinguSystems, Inc.
- Richard, Jean-Pierre J. 2011. "Does size matter?: The relationship between vocabulary breadth and depth". *Sophia International Review*, 107-119.
- Richards, Jack C. 1976. "The role of vocabulary teaching". *TESOL Quarterly* 10, 77-89.
- Richards Jack C. 2006. *Communicative Language Teaching Today*. Cambridge: Cambridge University Press.
- Richards, Lyn. 2005. *Handling Qualitative Data: A practical guide*. London: Sage.
- Robinson, Pauline. 1980. *ESP (English for specific purposes)*. Oxford: Pergamon Press.
- Roche, Thomas; Harrington, Michael. 2013. "Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting". *Language Testing in Asia* 3(12), 1-13.

- Roessingh, Hetty. 2016. "Academic Language in K-12: What is it, how is it learned, and how can we measure it?". *BC TEAL Journal* 1(1), 67-81.
- Roghani, Shadan; Milton, James. 2017. "Using category generation tasks to estimate productive vocabulary size in a foreign language". *TESOL International Journal* 12(1), 143- 159.
- Rubio, Olga G. 1997. "Ethnographic interview methods in researching language and education". In Hornberger, Nancy; Corson, David (eds.). *Encyclopedia of Language and Education, 8, Research methods in language education*. Dordrecht: Kluwer, 153–163.
- Sawaki, Yasuyo. 2013. "Structural equation modelling in language assessment". In Chapelle, Charol A. (ed.). *The Encyclopaedia of Applied Linguistics*. London: Blackwell Publishing, 1-6.
- Schermelleh-Engel, Karin; Moosbrugger, Helfried; Müller, Hans. 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, Norbert. 1998. "Tracking the incremental acquisition of second language vocabulary: A longitudinal study". *Language Learning* 48, 281-317.
- Schmitt, Norbert. 2010. *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.
- Schmitt, Norbert. 2019. "Understanding vocabulary acquisition, instruction, and assessment: A research agenda". *Language Teaching* 52, 261-274.
- Schmitt, Norbert; Carter, Ronald. 2004. "Formulaic sequences in action: An introduction". In Schmitt, Norbert (ed). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins, 1-22.
- Schmitt, Norbert; Meara, Paul. 1997. "Researching vocabulary through a word knowledge framework". *SSLA* 20, 17-36.
- Schmitt, Norbert; Nation, Paul; Kremmel, Benjamin. 2020. "Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation". *Language teaching* 53(1), 109-120.
- Schmitt, Norbert; Schmitt, Diane. 2014. "A reassessment of frequency and vocabulary size in L2 vocabulary teaching". *Language Teaching* 47(4), 484-503.
- Schmitt, Norbert; Schmitt, Diane; Clapham, Caroline. 2001. "Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test". *Language Testing* 18(1), 55-88.
- Schmitt, Norbert; Wun, Janice N., Garras, John. 2011. "The Word Associates Format: Validation evidence". *Language Testing* 28(1), 105-126.

- Schoonen, Rob; Verhallen, Marianne. 2008. "The assessment of deep word knowledge in young first and second language learners". *Language Testing* 25(2), 211-236.
- Schoonen, Rob. 2013. "Structural equation modelling". In Chapelle, Charol A. (ed.). *The Encyclopaedia of Applied Linguistics*. London: Blackwell Publishing, 1-9.
- Schreiber, James B.; Amaury, Nora; Stage, Frances K.; Barlow Elizabeth A.; King, James. 2006. "Reporting structural equation modelling and confirmatory factor analysis results: A review". *Journal of Educational Research* 99(6), 323-337.
- Schugurensky, Daniel. 2000. "The forms of informal learning: Towards a conceptualization of the field". *NALLWorking Report* 19, 1-8.
- Schur, Ellen. 2007. "Insights into the structure of L1 and L2 vocabulary networks: Intimidations of small worlds". In Daller, Helmut; Milton, James; Treffers-Daller; Jeanine (eds.). *Modelling and assessing vocabulary knowledge*. Cambridge: CUP, 182-203.
- Schurz, Alexandra; Coumel, Marion; Hüttner, Julia. 2022. "Accuracy and fluency teaching and the role of Extramural English: A tale of three countries". *Languages (Basel)* 7(35), 1-15. DOI: 10.3390/languages7010035
- Schuth, Elisabeth; Köhne, Judith; Weinert, Sabine. 2017. "The influence of academic vocabulary knowledge on school performance". *Learning and Instruction* 49, 157-165.
- Schwarz, Marlene. 2020. "Beyond the walls: A mixed methods study of teenagers' extramural English practices and their vocabulary knowledge". PhD thesis, University of Vienna.
- Seidlhofer, Barbara. 2005. "English as a lingua franca". *ELT Journal* 59(4), 339-341.
- Sercu, Lies; Peters Elke. 2002. "Learning e-learning – a comprehensive investigation of course developers' and language teacher trainees' views regarding the usefulness and effectiveness of a multimedia self-tuition course". *ReCALL* 14(1), 32-46.
- Shabani, Gholamhossein; Rahimy, Ramin. 2020. "An investigation of the effectiveness of enhancing input through glossing and skewing techniques on knowledge of lexical collocation". *Journal of Psycholinguistic Research* 49(6), 1011-1025.
- Shen, Li-Bi; Tseng, Ching-Ya; Kuo, Shu-Wei; Su, Ying-Ju; Chen, Ming-Yuan. 2005. "A preliminary study of college students' out-of-class English learning activities". *CHIA-NAN ANNUAL BULLETIN* 31, 464-475.
- Shin, Dongkwang; Nation, Paul I.S. 2008. "Beyond single words: the most frequent collocations in spoken English". *ELT Journal* 62(4), 339-348.
- Simpson-Vlach, Rita; Ellis, Nick C. 2010. "An academic formulas list: New methods in phraseology research". *Applied Linguistics* 31, 487-512.

- Sinclair, John. 1966. "Beginning the study of lexis". In Bazell, Charles E.; Catford, John C.; Halliday, Michael A.K.; Robins, Robert H. (eds.). *In memory of J.R. Firth*. London: Longmans, 410-430.
- Sinclair, John. 1987. "Collocation: A progress report". In Steele, Ross; Threadgold, Terry (eds.). *Language Topics: An International Collection of Papers by Colleagues, Students and Admirers of Professor Michael Halliday to Honour Him on His Retirement, Vol. II*. Amsterdam: John Benjamins Publishing Company, 319–331.
- Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, John. 2003. *Collins COBUILD advanced learner's English dictionary*. Glasgow: HarperCollins.
- Sinclair, John. 2004. "New evidence, new priorities, new attitudes". In Sinclair, John (ed.). *Vocabulary studies in first and second language acquisition*. Basingstoke: Palgrave, 271-299.
- Singson, Maria; Mahony Diana; Mann Virginia. 2000. "The relation between reading ability and morphological skills: Evidence from derivational suffixes". *Reading and Writing: An Interdisciplinary Journal* 12, 219-252.
- Siyanova Anna; Schmitt, Norbert. 2007. "Native and nonnative use of multi-word vs. one-word verbs". *IRAL* 45(2), 119-139.
- Skjelde, Kimberly; Coxhead, Averil. 2020. "Mind the gap: Academic vocabulary knowledge as a predictor of English grades". *Acta Didactica Norden* 14(3), 1-20.
- Socket, Geoffrey. 2013. "Understanding the online informal learning of English as a complex dynamic system: an emic approach". *ReCALL* 25(1), 48–62.
- Socket, Geoffrey. 2014. *The online informal learning of English*. New York: Palgrave Macmillan.
- Socket, Geoffrey; Toffoli, Denyze. 2012. "Beyond learner autonomy: A dynamic systems view of the informal learning of English in virtual online communities". *ReCALL* 24(2), 138–151.
- Söderqvist, Fredrik. 2019. "Perceptions of extramural English and English in the classroom: Swedish upper secondary students' writing, reading, listening and speaking skills". BA thesis, Dalarna University.
- Söderström, Filippa; Thorén, Jakob; Karlsson, Monika. 2022. "Pun intended: The use of word play to teach polysemous words in English as a second language". *International Journal of Language Studies* 16(2), 95-116.
- Sonbul, Suhad; Schmitt, Norbert. 2013. "Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions." *Language Learning* 63(1), 121-159.

- Sonbul, Suhad; El-Dakhs, Dina A.S.; Al-Otaibi, Hind. 2020. "Productive versus receptive L2 knowledge of polysemous phrasal verbs: A comparison of determining factors". *System* 95, 1-13. <https://doi.org/10.1016/j.system.2020.102361>
- Spratt, Mary; Humphreys, Gillian; Chan, Victoria. 2002. "Autonomy and motivation: Which comes first?". *Language Teaching Research* 6(3), 245–266.
- Staehr, Lars S. 2008. "Vocabulary size and the skills of listening, reading and writing". *Language Learning Journal* 36(2), 139-152.
- Staehr, Lars S. 2009. "Vocabulary knowledge and advanced listening comprehension in English as a foreign language". *SSLA* 31, 577-607.
- Stevens, James P. 2009. *Applied multivariate statistics for the social sciences* (5th edition). New York, NY: Taylor & Francis.
- Stewart, Jeffrey; Stoeckel, Tim; McLean, Stuart; Nation, Paul; Pinchbeck, Geoffrey G. 2021. "What the research shows about written receptive vocabulary testing". *Studies in Second Language Acquisition* 43, 462-471. doi:10.1017/S0272263121000437
- Stoeckel, Tim; Bennett, Phil; McLean, Stuart. 2016. "Is 'I don't know' a viable answer choice on the vocabulary size test?" *TESOL quarterly* 50(4), 965-975.
- Stoeckel, Tim; Ishii, Tomoko; Bennett, Phil. 2021. "Is the lemma more appropriate than the flemma as a word counting unit?". *Applied Linguistics* 41(4), 601-606. <https://doi.org/10.1093/applin/amy059>
- Stoeckel, Tim; McLean, Stuart; Nation, Paul I.S. 2021. "Limitations of size and levels tests of written receptive vocabulary knowledge". *Studies in Second Language Acquisition* 43, 181-203. doi:10.1017/S027226312000025X
- Stubbe, Raymond. 2012. "Do pseudoword false alarm rates and overestimation rates in yes-no vocabulary tests change with Japanese university students' ability levels?". *Language Testing* 29(4), 471–488. doi:10.1177/0265532211433033
- Stubbe, Raymond; Stewart, Jeffrey. 2012. "Optimizing scoring formulas for yes/no vocabulary checklists using linear models". *Shiken Research Bulletin* 16(2), 2–7. Retrieved from <http://teval.jalt.org/node/12>
- Suh, Jae-Suk; Wasanasomsithi, Punchalee; Short, Stephen; Majid, Norazman Abdul. 1999. "Out of class learning experiences and students' perceptions of their impact on English conversation skills". Retrieved from: <http://eric.ed.gov/?id=ED433715> (13.11.2020).
- Sullivan, Jennifer F. 2006. "Developing knowledge of polysemous vocabulary". PhD thesis, University of Waterloo.

- Sun, He; Steinkrauss, Rasmus; Wieling, Martijn; de Bot, Sun, Kees. 2018. "Individual differences in very young Chinese children's English vocabulary". *International Journal of Bilingual Education and Bilingualism* 21(4), 405-425.
- Sundqvist, Pia. 2009. "Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary". PhD thesis, Karlstad University.
- Sundqvist, Pia. 2013a. "Categorization of digital games in English language learning studies: Introducing the SSI model". In Bradley, Linda; Thouësny, Sylvie (eds.). *20 Years of EUROCALL: Learning from the past, looking to the future (2013 EUROCALL Conference Proceedings)*. Dublin: Research-publishing.net, 231–237.
- Sundqvist, Pia. 2013b. "The SSI Model: Categorization of digital games in EFL studies". *The European Journal of Applied Linguistics and EFL*, 89-104.
- Sundqvist, Pia. 2019. "Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary". *Language Learning & Technology* 23(1), 87–113.
- Sundqvist, Pia. 2022. "Media influence and identity. Young Swedish L2 English learners and changed self-perceptions". *YouthMediaLife-presentation*, University of Vienna, Austria, 13 October 2022.
- Sundqvist, Pia; Olin-Scheller, Christina. 2013. "Classroom vs. extramural English: Teachers dealing with demotivation". *Language and Linguistics Compass* 7(6), 329–338.
- Sundqvist, Pia; Sylvén, Liss K. 2012. "World of VocCraft: Computer games and Swedish learners' L2 vocabulary". *Digital games in language learning and teaching*. Basingstoke: Palgrave Macmillan, 189-208.
- Sundqvist, Pia; Sylvén, Liss K. 2014. "Language-related computer use: Focus on young L2 English learners in Sweden". *ReCALL* 26(1), 3–20.
- Sundqvist, Pia; Sylvén, Liss K. 2016. *Extramural English in teaching and learning: From theory and research to practice*. London: Palgrave Macmillan.
- Sundqvist, Pia; Uztosun, Serkan. (under review). "Extramural English in Scandinavia and Asia: Scale development, learner engagement, and perceived speaking ability".
- Sundqvist, Pia; Wikström, Peter. 2015. "Out-of-school digital gameplay and in-school L2 English vocabulary outcomes". *System* 51, 65–76.
- Sylvén, Liss K. 2006. "How is extramural exposure to English among Swedish school students used in the CLIL classroom?". *Vienna English Working Papers (VIEWS)* 15(3), 47–53.
- Sylvén, Liss K. 2018. "L2/FL English teaching and the challenge of the YouTube generation". Paper presented at *Exploring Language Education (ELE)*, Stockholm, 18-20 June. 2018

- Sylvén, Liss K. 2019a. "Extramural English". In Sylvén, Liss K. (ed.). *Investigating content and language integrated learning insights from Swedish high schools*. Bristol: Multilingual Matters, 152-167.
- Sylvén, Liss K. 2019b. "Looking at language through a camera lens". In Kalaja, Puala; Melo-Pfeifer, Sílvia (eds.). *Visualising multilingual lives: More than words*. Multilingual Matters: Bristol, 115-132.
- Sylvén, Liss K. 2021. "Gaming as a gateway to L2 English learning: A case study of a young L1 Swedish boy". In Peterson, Mark; Yamazaki, Kasumi; Thomas, Michael (eds.). *Digital games and language learning: theory, development and implementation*. London: Bloomsbury Academic, 213-236.
- Sylvén Liss K.; Löwenadler, John. 2022. "Let's play videos and L2 academic vocabulary". In Peterson, Mark; Jabbari, Nasser (eds.). *Digital Games in Language Learning: Case studies and applications*. London: Routledge, 93-108.
- Sylvén, Liss K.; Ohlander, Sölve. 2014. "The CLISS Project: Receptive vocabulary in CLIL versus non-CLIL groups". *Moderna spark*, 2014(2), 80-114.
- Sylvén, Liss K.; Sundqvist, Pia. 2012. "Gaming as extramural English L2 learning and L2 proficiency among young learners". *ReCALL* 24(3), 302–321.
- Sylvén, Liss K.; Sundqvist, Pia. 2016. "Validation of a test measuring young learners' general L2 English vocabulary knowledge." *Novitas-ROYAL* 10(1), 1-23.
- Sylvén, Liss K.; Ohlander, Sölve. 2014. "The CLISS Project: Receptive vocabulary in CLIL versus non-CLIL groups". *Moderna spark*, 2014(2), 80-114.
- Szabo, Csaba; Stickler, Ursula; Adinolfi, Lina. 2021. "Predicting the academic achievement of multilingual students of English through vocabulary testing". *Journal of Bilingual Education and Bilingualism* 24(10), 1531-1542, DOI: 10.1080/13670050.2020.1814196
- Szudarski, Paweł; Carter, Ronald. 2016. "The role of input flood and input enhancement in EFL learners' acquisition of collocations". *International Journal of Applied Linguistics* 26(2), 245-265.
- Takumi, Uchiyara; Kazuya, Saito. 2019. "Exploring the relationship between productive vocabulary knowledge and second language oral ability". *Language Learning Journal* 47(1), 64-75.
- Talakoob, Fahimeh; Koosha, Mansour. 2017. "Productive and receptive collocational knowledge of Iranian EFL learners at different proficiency levels". *International Journal of Applied Linguistics & English Literature* 6(7), 11-16.
- Tahmasbi, Maryam; Farvardin, Mohammad T. 2017. "Probing the effects of task types on EFL learners' receptive and productive vocabulary knowledge: The case of

- involvement load hypothesis". *SAGE Open* 7(3), 1-10.
<https://doi.org/10.1177/2158244017730596>
- Tannenbaum, Kendra R.; Torgesen, Joseph K.; Wagner, Richard K. 2006. "Relationships between word knowledge and reading comprehension in third-grade children". *Scientific Studies of Reading* 10, 381–98.
- Tashakkori, Abbas. 2006. *Mixed methodology: combining qualitative and quantitative approaches*. Thousand Oaks: SAGE.
- Thorndike, Edward L.; Lorge Irving. 1944. *The Teacher's Word Book of 30, 000 Words*. New York: Teachers College, Columbia University.
- Toffoli, Denyze; Sockett, Geoffrey. 2016. "University teachers' perceptions of Online Informal Learning of English (OILE)". *Computer Assisted Language Learning* 28(1), 7-21.
- Townsend, Dianna; Bear, Donald; Templeton, Shane; Burton, Amy. 2016. "The implications of adolescents' academic word knowledge for achievement and instruction". *Reading Psychology*, 1-30.
- Townsend, Dianna; Filippini, Alexis; Collins, Penelope; Biancarosa, Gina. 2012. "Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students". *The Elementary School Journal* 112(3), 497-518.
- Tran, Mai P. 2017. "Exploring young learners' informal learning of English language: A comparative study on the perspectives of 11-13-year-old pupils in Finland and Vietnam". MA thesis, University of Eastern Finland.
- Trinder, Ruth. 2015. "Online informal learning of English: How students use technology to supplement classes: International Conference ICT for Language Learning" Retrieved from: <http://conference.pixel-online.net/ICT4LL/files/ict4ll/ed0008/FP/2161-ICL1391-FP-ICT4LL8.pdf> (18 Nov. 2020).
- Trinder, Ruth. 2017. "Informal and deliberate learning with new technologies". *ELT Journal* 71(4), 401–412.
- Tschirner, Erwin. 2004. "Breadth of vocabulary and advanced English study: An empirical investigation". *Electronic Journal of Foreign Language Teaching* 1(1), 26-38.
- Tyler, Andrea; Nagy William. 1990. "Use of derivational morphology during reading". *Cognition* 36, 17-34.
- Uchihara, Tamkui; Saito, Kazuya. 2019. "Exploring the relationship between productive vocabulary knowledge and second language oral ability". *The Language Learning Journal* 47 (1), 65-75. ISSN 0957-1736.

- Ullmann, Jodie B. 2007. "Structural equation modelling". In Tabachnick, Barbara; Fidell, Linda S. (eds.). *Using multivariate statistics*. Boston: Pearson/Allyn and Bacon, 676-780.
- Vafaei, Payman; Suzuki, Yuichi. 2020. "The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability". *Studies in Second Language Acquisition* 42, 383-410.
- Vagh, Shaher B.; Pan, Barbara A.; Mancilla-Martinez, Jeannette. 2009. "Measuring growth in bilingual and monolingual children's English productive vocabulary development: The utility of combining parent and teacher report". *Child Development* 80(5), 1545-1563.
- Van Zeeland, Hilde; Schmitt, Norbert. 2013. "Incidental vocabulary acquisition through L2 listening: A dimensions approach". *System* 41, 609-624.
- Verspoor, Marjolijn H.; De Bot, Kees; Van Rein, Eva. 2011. "English as a foreign language: The role of out-of-school language input". In De Houwer, Annick; Wilton, Antje (eds.). *English in Europe today: Sociocultural and educational perspectives*. Amsterdam: John Benjamins, 147-166.
- Verspoor, Marjolijn H.; Lowie, Wander. 2003. "Making sense of polysemous words". *Language Learning* 53, 547-586.
- Vienna English Language Test (VELT). 2022. Vienna: University of Vienna. Retrieved from: <https://spl-anglistik.univie.ac.at/study/beginning-your-studies/velt-1st-term/> (31.12.22).
- Vilkaitė, Laura; Schmitt, Norbert. 2019. "Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations?". *Applied Linguistics* 40(2), 329-354.
- Vilkaitė-Lozdienė, Laura; Schmitt, Norbert. 2020. "Frequency as a guide for vocabulary usefulness: High-, mid- and low- frequency words". In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 81-96.
- Vives Boix, G. 1995. "The development of a measure of lexical organization: The Association Vocabulary Test". Unpublished PhD thesis. University of Wales, Swansea.
- Von Eye, Alexander; Spiel, Christiane; Wagner, Petra. 2003. "Structural equation modelling in developmental research concepts and applications". *Methods of Psychological Research Online* 8(2), 75-112.
- Vongpumivitch, Vipavee; Huang, Ju-yu; Chang, Richard Y. 2009. "Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers". *English for Specific Purposes* 28(1), 33-41.

- Wang, Zhaogang. 2014. "A correlation analysis on the depth and breadth of ESL learners' vocabulary knowledge and their overall linguistic competence". *Theory and Practice in Language Studies* 4(12), 2460-2465.
- Wang, Jing; Liang, Shao-lan; Ge, Guang-chun. 2008. "Establishment of a medical academic word list". *English for Specific Purposes* 27, 442-458.
- Ward, Jeremy. 2009. "A basic engineering English word list for less proficient foundation engineering undergraduates". *English for Specific Purposes* 28, 170-182.
- Warnby, Marcus. 2022. "Receptive academic vocabulary knowledge and extramural English involvement: is there a correlation?". *ITL - International Journal of Applied Linguistics* 173(1), 120-152.
- Watson-Todd, Richard W. 2017. "An opaque engineering word list: Which words should a teacher focus on?". *English for Specific Purposes* 45, 31-39.
- Webb, Stuart. 2005. "Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge". *Studies in Second Language Acquisition* 27, 33-52.
- Webb, Stuart. 2007. "The effects of repetition on vocabulary knowledge". *Applied Linguistics* 28, 46-65.
- Webb, Stuart. 2008. "Receptive and productive vocabulary sizes of L2 learners". *SSLA* 30, 79-95.
- Webb, Stuart. 2009. "The effects of receptive and productive learning of word pairs on vocabulary knowledge". *RELC* 40(3), 360-376.
- Webb, Stuart; Chang, Anna. 2012. "Second language vocabulary growth". *RELC* 43(1), 113-126.
- Webb, Stuart; Kagimoto, Eve. 2009. "The effects of vocabulary learning on collocation and meaning". *TESOL Quarterly* 43(1), 55-77.
- Webb, Stuart; Newton, Jonathan; Chang, Anna. 2013. "Incidental Learning of Collocation". *Language Learning* 63(1), 91-120.
- Webb, Stuart; Rodgers, Michael P.H. 2009. "The lexical coverage of movies". *Applied Linguistics* 30(3), 407-427.
- Webb, Stuart; Sasao, Yosuke; Ballance, Oliver. 2017. "The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT". *ITL – International Journal of Applied Linguistics* 168(1), 33–69.
- West, Michael. 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman.

- Wichman, Søren. 2015. "Statistical observations on implicational (verb) hierarchies." In Malchukov, Andrej; Comrie, Bernard (eds.). *Introducing the framework, and case studies from Africa and Eurasia*. Berlin: De Gruyter, 155-181.
- Wilkinson, Sharon. 2002. "The omnipresent classroom during summer study abroad: American students in conversation with their French hosts". *The Modern language journal* 86(2), 157-173.
- Winke, Paula; Gass, Susan; Sydorenko, Tetyana. 2010. "The effects of captioning videos used for foreign language listening activities". *Language Learning and Technology* 14(1), 65–86. <http://dx.doi.org/10.125/44203>
- Wolter, Brent. 2001. "Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model". *Studies in Second Language Acquisition* 23(1), 41–69.
- Wolter, Brent; Gyllstad, Henrik. 2013. "Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations". *Studies in Second Language Acquisition* 35, 451-482.
- Wood, David. 2019. "Classifying and identifying formulaic language". In Webb, Stuart (ed.). *The Routledge Handbook of Vocabulary Studies*. London: Routledge, 30-45.
- Woodward-Kron, Robyn. 2008. "More than just jargon: The nature and role of specialist language in learning disciplinary knowledge". *Journal of English for Academic Purposes* 7, 234-249.
- Wray, Alison. 1992. *The focusing hypothesis: The theory of left hemisphere lateralized language re-examined*. Amsterdam: John Benjamins.
- Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Xue, Guoyi; Nation Paul I.S. 1984. "A university word list". *Language learning and communication* 3, 215-299.
- Yamamoto, Yuka. 2014. "Multidimensional vocabulary acquisition through deliberate vocabulary list learning". *System* 42, 232-243.
- Yang, Huizhong. 1986. "A new technique for identifying scientific/ technical terms and describing science texts". *Literary and Linguistic Computing* 1, 93-103.
- Yang, Yingli; Sun Ya; Chang, Pengyun; Li, Yumo. 2019. "Exploring the relationship between language aptitude, vocabulary size, and EFL graduate students' L2 writing performance". *TESOL Quarterly* 53(3), 845-856.
- Yap, Set-lee S. 1998. "Out-of-class use of English by secondary school students in a Hong Kong Anglo-Chinese school". MA thesis, University of Hong Kong.

- Yi, Youngjoo. 2005. "Asian adolescents' out-of-school encounters with English and Korean literacy". *Journal of Asian Pacific Communication* 15(1), 57-77.
- Zang, Xian; Lu, Xiaofei. 2015. "The relationship between vocabulary learning strategies and breadth and depth of vocabulary knowledge". *The Modern Language Journal* 99(4), 740-753.
- Zhang, Xiapeng; Wen, Ju. 2019. "Exploring multiple constraints on second language development of English polysemous phrasal verbs". *Applied Psycholinguistics* 40, 1073–1101.
- Zhong, Hua F. 2018. "The relationship between receptive and productive vocabulary knowledge: a perspective from vocabulary use in sentence writing". *The Language Learning Journal* 46(4), 357-370.
- Zichtl, Carolina. 2017. "A study of receptive and productive lexical knowledge of Austrian EFL students in year 8, 10, and 12". MA thesis, University of Vienna.
- Zuraina, Ali; Mukundan, Jayakaran; Baki, Roselan; Ayub, Ahmad F.M. 2012. "Second language learners' attitudes towards the methods of learning vocabulary". *English Language Teaching* 5(4), 24-36.

Digital Appendix

All appendices are permanently stored in Phaidra, an online repository providing permanent and safe storage of digital files, run by the University of Vienna. This is the link to the complete collection connected to this thesis:

<https://phaidra.univie.ac.at/o:1632884>

Deutsche Zusammenfassung

Aufbauend auf bereits existente Forschung (z.B. Lee, 2019; Lee & Dressman, 2018; Lee & Sylvén, 2021; Sundqvist & Sylvén, 2014), welche die Wichtigkeit von Extramural English (EE) für den Erwerb von Zweitsprachenvokabular hervorhebt, beschäftigt sich diese Studie eingehend mit zwei Forschungsbereichen: Einerseits mit der Kenntnis akademischer englischer Verben und andererseits mit deren potenzieller Verbindung zu Extramural English Aktivitäten von 147 Englischsstudenten in Wien. Angelehnt an Nation's (2013) Taxonomie des Wortwissens, konzentriert sich die Studie auf folgende vier Komponenten: form-meaning recognition, form-meaning recall, multiple meaning recognition und collocation recognition. Dabei lag der Fokus auf akademischen Verben verschiedener Häufigkeitslevels.

Für diesen Zweck wurden vier Tests, einer für jede Komponente des genannten Wortwissens entwickelt, welche dieselben akademischen Verben beinhalten. Das zweite Konzept, Extramural English, breit definiert als Englischkontakt außerhalb der Wände des Klassenzimmers (frei übersetzt nach Sundqvist & Sylvén, 2016, p. 6)., wird mit Hilfe eines selbstentwickelten Fragebogens erhoben. Des Weiteren erlaubt die Studienlaufzeit von über einem Jahr Einblicke in die Langzeitentwicklung beider Ebenen und deren Verbindungen.

Obwohl akademisches Verbwissen und EE Verhalten im tertiären Bildungssektor bereits früher untersucht wurden (z.B. Busby, 2020; Lee & Sylvén, 2021; Sundqvist & Sylvén, 2014), erlaubt diese Studie neue Erkenntnisse in den Verbindungen zwischen Komponenten des Wortwissens, und bemerkenswert stabile EE Beteiligung über die Zeit gezeigt werden. Des Weiteren wurden lineare und multiple lineare Regressionsmodelle eingesetzt, um den Einfluss von EE Beteiligung auf das akademische Verbwissen zu untersuchen. Diese zeigten, dass bestimmte EE Aktivitäten akademisches Verb Wissen signifikant beeinflussen.