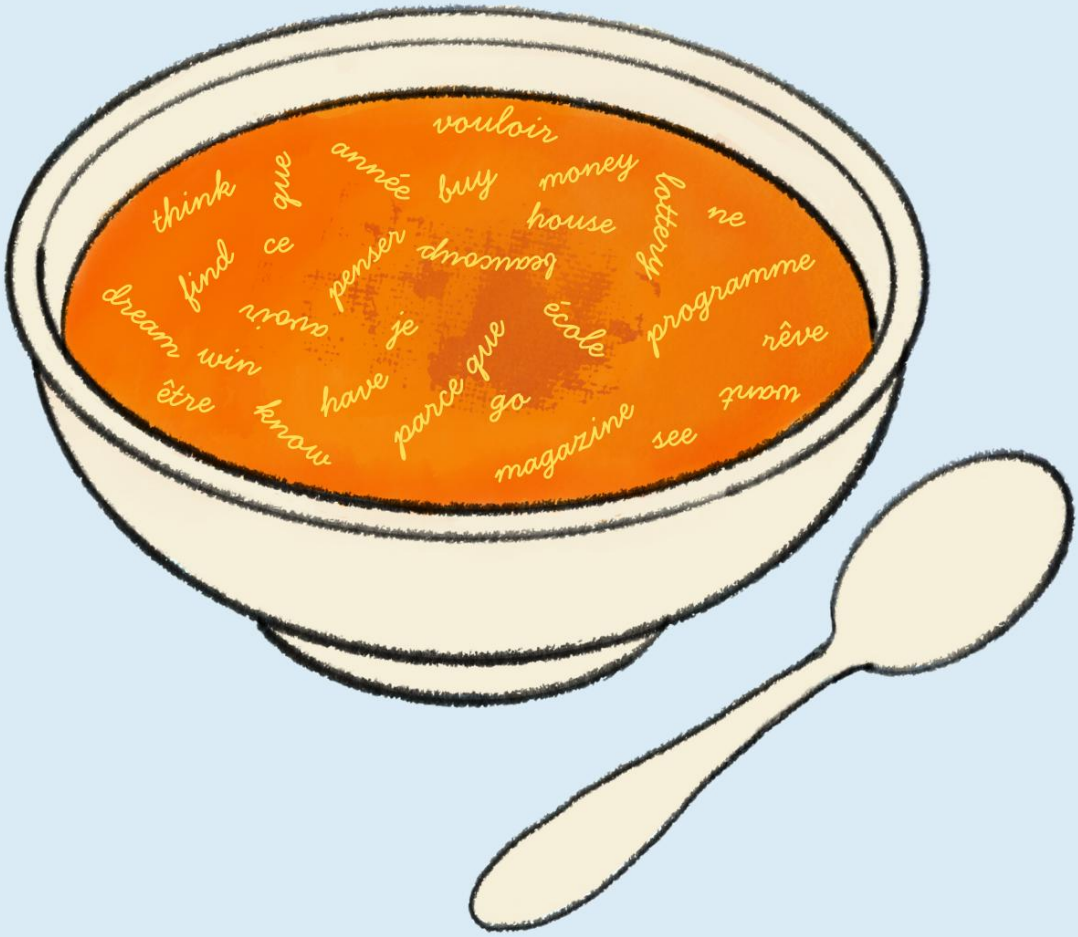
What's the word?

A multi-dimensional investigation of vocabulary knowledge and vocabulary use by second language learners of English and French



Eva Caltabellotta

KU Leuven – Faculty of Arts

Campus Antwerp



What's the word?

A multi-dimensional investigation of vocabulary knowledge and vocabulary use by second language learners of English and French

Doctoral dissertation submitted in order to obtain the academic degree of

Doctor (PhD) of Linguistics

by

Eva Caltabellotta

Supervisor: Elke Peters
Co-supervisor: Elke Van Steendam
Co-supervisor: Ann-Sophie Noreillie

16 December 2025

Examination committee:

Supervisor:	Prof. Dr. Elke Peters
Co-supervisors:	Prof. Dr. Elke Van Steendam
	Prof. Dr. Ann-Sophie Noreillie
External members:	Prof. Dr. Rob Schoonen
	Prof. Dr. Beatriz González-Fernández
Internal members:	Prof. Dr. Kris Van den Branden
	Prof. Dr. Morgane Jourdain
Chair:	Prof. Dr. Geert Brône

Cover design: Merel Smet, 2025. 'Letter(en)soep'. Words selected from the learner corpus collected for this dissertation.

Funding: The studies included in this dissertation were funded by KU Leuven Internal Funds (IMP/20/008).

Ethical approval: This research was reviewed and approved by the Social and Societal Ethics Committee (SMEC) at KU Leuven, approval number G-2020-2902-R3(MIN).

Use of AI: Generative AI was used to edit programming code for Python and R in the empirical studies (OpenAI, 2024).

Words are, in my not-so-humble opinion,
our most inexhaustible source of magic.

Albus Dumbledore,
Harry Potter and the Deathly Hallows

Table of contents

Table of contents I
List of tables VII
List of figures IX
Acknowledgements | Dankwoord X
1. Introduction 1
1.1. Rationale and research aims 3
1.2. Structure of the thesis 7
1.3. Key terminology 8
1.4. Bibliographical information 11
2. Background literature 13
PART I: Vocabulary knowledge 13
2.1. Defining vocabulary knowledge 13
2.1.1. What is a word? 13
2.1.2. Knowing a word 16
2.1.3. Dimensions of mastery 18
2.2. Measuring vocabulary knowledge 20
2.2.1. Vocabulary test formats 20
2.2.2. Selected-response vocabulary tests 21
2.2.3. Constructed-response vocabulary tests 25
2.3. Measuring vocabulary use 28
2.3.1. Indices of lexical richness 28
2.3.2. Pre-processing learner texts 37
2.3.3. The role of the topic and instructions 39
PART II: Vocabulary and skills 41
2.4. Vocabulary knowledge and skills 41
2.4.1. Reading and listening 41
2.4.2. Speaking and writing 43
2.5. Vocabulary use and skills 44
2.5.1. Lexical diversity 45
2.5.2. Lexical sophistication 46
PART III: Vocabulary learning 48
2.6. Theoretical accounts of vocabulary learning 48

2.6.1. Usage-based learning 48
2.6.2. Input hypothesis 50
2.7. Instructed learning 52
2.7.1. Instruction and L2 vocabulary learning 52
2.7.2. Word-focused instruction 55
2.8. Out-of-school learning 59
2.8.1. Defining out-of-school contact 59
2.8.2. Out-of-school contact and L2 vocabulary learning 61
2.8.3. Out-of-school activities and L2 vocabulary learning 63
Conclusion 66
3. Methodology 67
3.1. Design 67
3.2. Research context and participants 67
3.3. Data collection instruments 71
3.3.1. Pilot study 71
3.3.2. Final instruments 77
3.4. Data collection procedure 81
3.5. Scoring and lexical analyses 83
3.5.1. Productive vocabulary knowledge 83
3.5.2. Vocabulary use 84
Summary 92
4. Vocabulary knowledge and vocabulary use in writing: A cross-sectional study 93
Abstract 93
4.1. Introduction 94
4.2. Literature review 94
4.2.1. L2 vocabulary knowledge 94
4.2.2. L2 vocabulary use 95
4.2.3. Vocabulary knowledge and use 97
4.3. Research questions 98
4.4. Method 99
4.4.1. Design and procedure 99
4.4.2. Participants 99
4.4.3. Instruments 100

4.4.4. Scoring 101
4.4.5. Data analysis 102
4.5. Results 103
4.5.1. Cross-sectional differences in vocabulary knowledge and use 103
4.5.2. L2 and L1 vocabulary use 105
4.5.3. Vocabulary knowledge and grade as predictors of vocabulary use 106
4.6. Discussion 108
4.6.1. Cross-sectional differences in vocabulary knowledge and use 108
4.6.2. Vocabulary knowledge and grade as predictors of vocabulary use 110
4.6.3. Pedagogical implications 111
4.6.4. Limitations and future directions 112
4.7. Conclusion 113
5. The role of out-of-school contact and instruction in productive vocabulary knowledge and vocabulary use 115
Abstract 115
5.1. Introduction 116
5.2. Literature review 117
5.2.1. Out-of-school L2 contact and vocabulary knowledge 117
5.2.2. Out-of-school L2 contact and vocabulary use in writing 119
5.2.3. Formal language instruction 120
5.3. Rationale and research questions 121
5.4. Methodology 122
5.4.1. Research context and participants 122
5.4.2. Design and procedure 122
5.4.3. Instruments 123
5.4.4. Measures of vocabulary knowledge and vocabulary use 124
5.4.5. Statistical analyses 125
5.5. Results 127
5.5.1. Descriptive findings 127
5.5.2. RQ1: The relationship between out-of-school contact, instruction, gender and vocabulary measures 130

5.5.3. RQ2: The role of out-of-school L2 activities in vocabulary knowledge and use 133
5.6. Discussion 135
5.6.1. The relationship between out-of-school contact, instruction and vocabulary measures 135
5.6.2. The role of out-of-school activities in vocabulary knowledge and use 137
5.6.3. Limitations and future directions 138
5.7. Conclusion 139
6. A methodological perspective on measuring vocabulary use 141
Abstract 141
6.1. Introduction 142
6.2. Literature review 143
6.2.1. Vocabulary use in writing 143
6.2.2. Lemmatization 145
6.2.3. Word inclusion 146
6.3. Rationale and research questions 147
6.4. Methodology 148
6.4.1. Design and participants 148
6.4.2. Instruments 148
6.4.3. Lexical analyses 149
6.4.4. Statistical analyses 152
6.5. Results 153
6.5.1. Differences in lexical values between operationalizations 153
6.5.2. Relationship between vocabulary knowledge and each operationalization 157
6.6. Discussion 160
6.6.1. The effect of lemmatization and word inclusion 160
6.6.2. The relationship with productive vocabulary knowledge 161
6.6.3. Limitations and future directions 162
6.7. Conclusion 163
7. Conclusions 165
7.1. Summary of the three studies 165
7.1.1. Study 1 167

7.1.2. Study 2 168
7.1.3. Study 3 169
7.2. Theoretical contribution 170
7.2.1. Cross-sectional development 170
7.2.2. The relationship between productive vocabulary knowledge and vocabulary use 174
7.2.3. The role of out-of-school contact and formal instruction 175
7.2.4. Cross-linguistic differences 178
7.3. Pedagogical contribution 179
7.3.1. Selection of vocabulary 180
7.3.2. Vocabulary activities 182
7.3.3. Cross-linguistic differences 184
7.4. Methodological contribution 185
7.4.1. Sample and research design 185
7.4.2. Writing tasks 186
7.4.3. Lexical diversity and lexical sophistication indices 188
7.4.4. Pre-processing of learner texts 189
7.4.5. Productive vocabulary test 192
7.4.6. Questionnaire: out-of-school contact estimates 193
7.5. Limitations and future directions 194
7.5.1. Questionnaire 194
7.5.2. Writing tasks 195
7.5.3. Scoring and lexical analyses 196
7.5.4. Generalizability 197
7.6. Conclusion 197
Final remarks 199
References 203
Appendices 243
A1. Productive vocabulary tests 243
A1.1. Productive VocabLab test – English 243
A1.2. Productive VocabLab test – French 247
A2. Writing tasks 251
A2.1. 'Prize' prompts 251
A2.2. 'Complaint' prompts 252

A3. Questionnaires 253
A3.1. Questionnaire – original 253
A3.2. Questionnaire – translation 258

List of tables

Table 1.1. Detailed research questions that guided the three studies 7

Table 2.1. Illustration of the different counting units for the headwords *run* (English) and *marche* (French) 15

Table 2.2. Illustration of decreasing TTR with increased repetitions 32

Table 3.1. Characteristics of participants in the full sample (N = 423) 69

Table 3.2. Mean and median number of tokens produced on the writing tasks in the pilot study (n = 65) 76

Table 3.3. Example sentences from the productive VocabLab tests in English and French..... 78

Table 3.4. Overview of the writing prompts 79

Table 3.5. Example sentence demonstrating tokenization and POS tagging in spaCy 85

Table 3.6. Illustration of default and custom lemmatization in Study 3 ... 86

Table 3.7. Example of MATTR calculation based on lemmatized 5-word segments 88

Table 4.1. Descriptive results of vocabulary tests and writing output..... 104

Table 4.2. Mann-Whitney U-test results comparing grade 10 and grade 12 participants in terms of vocabulary knowledge, lexical diversity and lexical sophistication..... 104

Table 4.3. Mann-Whitney U-test results comparing grade 10 and grade 12 participants in terms of band-based lexical sophistication.....105

Table 4.4. Mann-Whitney U-test results comparing output by L2 and L1 writers..... 106

Table 4.5. Spearman correlation matrix of the vocabulary measures 106

Table 4.6. Mixed-effects models for English with form recall scores and grade as predictor variables 107

Table 4.7. Mixed-effects models for French with form recall scores and grade as predictor variables 108

Table 5.1. Means and standard deviations for learners' form recall knowledge (%), lexical diversity (MATTR), and count-based (ALF) and band-based lexical sophistication (LFP) 129

Table 5.2. Descriptive results for learners' length of instruction (in years), overall out-of-school contact and exposure to specific out-of-school activities, reported in hours per week.....129

Table 6.1. Illustration of default and custom lemmatization..... 151

Table 6.2. Descriptive findings for MATTR, ALF and LFP for all four operationalizations in English and French 153

Table 6.3. Spearman correlation matrix for the four operationalizations 153

Table 6.4. Results of linear mixed-effects models for MATTR, ALF and LFP in English and French 156

Table 6.5. Results of English GAMs with productive vocabulary knowledge as predictor and four operationalizations of lexical indices as dependent variables.....158

Table 6.6. Results of French GAMs with productive vocabulary knowledge as predictor and four operationalizations of lexical indices as dependent variables.....159

List of figures

Figure 1.1. Overview of the three studies 5

Figure 3.1. Example item from the section on out-of-school L2 contact on the French questionnaire 81

Figure 5.1. PLS-SEM results for L2 English 131

Figure 5.2. PLS-SEM results for L2 French 132

Figure 5.3. Plots visualizing the Generalized Additive Mixed Model results for English 134

Figure 5.4. Plots visualizing the Generalized Additive Mixed Model results for French 134

Figure 7.1. Overview of the three studies 166

Acknowledgements | Dankwoord

Four years of work – yet somehow it feels like it took three months and ten years at the same time.

While this dissertation may have been written by me, I am in no way the only contributor to the end product lying before you.

First, I would like to extend my gratitude to my **supervisors**, 'the Elkes' and Ann-Sophie. **Elke P**, thank you for giving me so many opportunities to grow, from the first year of my Bachelor program to the last year of my PhD. Thank you for believing in me, encouraging me and challenging me. Your support and guidance has helped me to become who I am today. **Elke VS** and **Ann-Sophie**, thank you for all of your feedback, support and encouragement. I very much appreciated your insights and your regular check-ins. I will surely miss receiving comments from the 'EP_EVS_ASN' tryad.

In the spirit of feedback: a genuine thank you to the members of my **doctoral supervisory committee**, Rob Schoonen and Eva Puimège. **Rob**, while our meetings may not have been numerous, I much appreciated your always constructive and valuable feedback. **Eva**, thank you for being my 'residuals plot expert'. Non-normality and heteroscedasticity are ever-present, but you taught me that sometimes, imperfect will have to do. I will, however, be double-checking my plots with you until forever.

I would also like to gratefully acknowledge the members of my **examination committee**: Prof. **Rob** Schoonen, Prof. **Beatriz** González-Fernández, Prof. **Kris** Van Den Branden, and Prof. **Morgane** Jourdain. Thank you for taking the time to read and evaluate my thesis, and for spending a day talking to each other about me. This may be the closest I ever get to being famous ☺. My sincerest thanks also goes to Prof. **Geert** Brône, who was so kind to be the chair of my defense. Your calm, soothing, and kind energy is probably exactly what I will need during my defense (I'll keep you posted on that afterwards).

I would not have enjoyed this PhD trajectory as much as I did without my dearest colleagues. Allereerst, **mijn lieve kantoorgenootjes**: Elisa, Marie, Joke, Matilde, en eerder ook Elien, Jolien (en heel even Kambiz). **Elisa en Marie**, wat was het fijn om samen met jullie de eindfase te beleven (maar uiteraard de andere jaren ook hé!). Blij dat we samen af en toe konden panikeren en tegen elkaar op konden klagen. **Joke**, dank je wel om niet alleen een leuke kantoorgenoot te zijn, maar ook een vriendin, een lunchdate, een sparringpartner over onderzoekskwesties, en een spinnenverwijderaar (al was die ene grote spin in de gang ook voor jou te

veel van het goede). Heel blij met jou als 'bur-o-verbuur' (snap je het?) in dit laatste jaar.

To **Sinem, Elle, Charlien, Ruben, Justyna, Fien, Dory** and the other colleagues on campus Antwerp, I loved having you around for some company in the office, regular chats and an occasional 'domibo' (although I firmly stand by my claim that nobody actually uses that term). Een speciaal bedankje voor de collega's van de eerste verdieping die mij een tijdje officieus geadopteerd hebben, in het bijzonder Eva P., Inez, Karen en Priscilla. **Eva P.** en **Inez**, dank jullie wel om onze campus zo warm en gezellig te maken. De beledigende grapjes die ik soms moest incasseren zal ik maar niet te serieus nemen. **Karen** en **Priscilla**, dank jullie wel om samen met Elke onze campus recht te houden. Jullie zijn ware superhelden. Bedankt ook aan de **ATP'ers** om keer op keer onze (soms domme) vragen te beantwoorden (**Ann, Wim, alle STOFF-collega's, de bib-collega's, Veronique en Shana**: ik kijk naar jullie). Aan de collega's van **Stuvo**, dank jullie wel voor de fijne samenwerkingen en contacten. Jullie zijn top.

Raf en **Inge D.**, een oprechte dankjewel om de keuze voor de opleiding Toegepaste Taalkunde zo gemakkelijk te maken. Mijn eerste ontmoeting met jullie tijdens de openlesdagen (nu al 10 jaar geleden!) heeft ervoor gezorgd dat ik mij meteen welkom en thuis voelde op campus Sint-Andries. Het is dankzij jullie dat ik de 'juiste' opleiding heb gekozen om dit pad überhaupt te kunnen bewandelen.

Outside of the Antwerp campus, thank you to the other ABAP **colleagues at campus Leuven, Brussel and Kortrijk**. It's been great having you around for the occasional chat or reception. I would especially like to thank my fellow **representatives in the ABAP Council** for your commitment. Een speciale shoutout naar **Melissa** en **Sinem**, jullie zijn echt fantastische voorzitters (geweest). Dank jullie wel om jullie zo in te zetten voor het ABAP en voor onze faculteit. **Karen** L. and **Eline** Z. (and by extension **Sylvia**), thank you for making our research unit a safe and welcoming space, in particular for junior researchers. **Lena** and **Kahn**, thank you for all of your initiatives to inform and support the ABAP, and thanks for reading a million grant applications every year. We much appreciate your feedback. I would also like to sincerely thank all of the (former) members of the **Vocab group**: Lieven, Elien, Eva P., Duy, Marion, Ann-Sophie, Jolien, Joke. Thank you for the feedback, and thank you for the support. I can't wait to see where life takes you all.

Speaking of feedback and support: **my dearest fellow researchers at the University of Oregon**, in particular Kris, Hakyung, Carissa, Carla and Karla. **Kris**, thanks for inviting me and for creating a welcoming space for me, thanks for your flexibility and availability, and all of your input and feedback. To the others, apart from the fact that I learned a bunch from you,

I loved exploring Oregon with you guys. **Hakyung** and **Carissa**, thank you for taking me to the stunning Silver Falls. **Carla** with a C, thank you for letting me stay at your apartment even though you barely knew me (you're crazy for that). I also loved our trip to the coast, thank you for showing me the Pacific Ocean from up close. **Karla** with a K, thanks for our little chats, food breaks and even a theater night. Thank you all for having me – I truly had a blast.

Mijn bijzondere dank gaat ook uit naar alle **leerkrachten en leerlingen** die hebben meegewerkt aan de dataverzameling. Bedankt, leerkrachten, om jullie zo flexibel op te stellen, om te helpen organiseren, en om enkele lesuren af te staan. Bedankt, leerlingen, om deel te nemen aan het onderzoek en een hele hoop taken af te leggen (de ene al met wat meer enthousiasme dan de andere ☺). Bovendien bedankt aan de **directieleden en secretariaatsleden** die hun steentje hebben bijgedragen aan de organisatie en het goede verloop van de dataverzameling.

Ook in mijn **privéleven** hoefde ik deze vier jaar niet alleen te trotseren. **Maurice**, dank u om mij uit te dagen, dank u om mij te inspireren, en vooral dank u om er te zijn. Het was een absoluut voorrecht om de afgelopen drie jaar samen met u de wereld te mogen ontdekken. **Peejee**, wat een geluk heb ik met u. Dank u wel om uzelf te zijn, en dank u wel vooral dat ik ook mijzelf mag zijn. We wisten het allebei al wel, die 'haat-liefderelatie' van in het middelbaar is altijd meer liefde dan haat geweest. Ik heb echt de beste beste vriend ooit, en ik ben heel dankbaar voor onze 'one-on-one date turns double date'-dagen met Merel. Over double dates gesproken, **Lieven en Silke**, dank jullie wel om altijd voor een paar uur rust te brengen in mijn hoofd. Sorry dat ik nog altijd niet kan onthouden wat er nu weer precies gebeurt in *The Lord of the Rings*... We zullen de films nog eens opnieuw moeten kijken! Op nog heel veel double dates in de toekomst. **Anessa**, merci om als klein 'prutske' in de kleuterklas al aan mijn zijde te staan. Nooit gedacht dat vriendschap voor het leven zo letterlijk kon zijn. Ik kan niet wachten om te zien hoe ge de rest van uw leven uitbouwt. **Elise**, mijn beste vriendje tijdens onze bacheloropleiding, en ook wel een beetje mijn rolmodel. Het was me een eer om op vrijdagavonden samen met u van campus Sint-Andries naar het station te mogen hollen, omdat we maar 22 minuten hadden om met ons hele hebben en houwen op de trein te raken. Ik ben heel dankbaar dat we elkaar niet uit het oog verloren zijn. En tot slot, **mijn lieve mama.** Dank u wel om er altijd te zijn voor mij, en niet alleen tijdens mijn doctoraat. 'Als ik later groot ben', wil ik zoals gij zijn. Dank u wel om er te zijn, dank u wel om uzelf te zijn. Gij zijt mijn rots, every step of the way. Ik ben ongelooflijk dankbaar dat gij mijn mamaatje zijt. ♡

1. Introduction

Writing is a central skill in learning a second or foreign language (L2). When writing a text in an L2, learners are faced with the difficult task of turning the ideas in their mind into words on a page. This can be quite the challenge, as learners are required to manage multiple demands at once: they have to generate ideas, organize these ideas coherently, and express the ideas accurately with sufficient nuance and complexity, all while drawing on the range of L2 words available in their mental lexicon. Research on L2 writing development has indicated that this presents a considerable challenge for learners (Manchón, 2011; Myles, 2022; Ortega, 2009; Schmitt, 2020). Yet, the ability to write effectively is essential in communicative contexts, which is why writing is a major component of large-scale standardized assessment and why it holds a place in virtually all language curricula (Knoch, 2021; Leow et al., 2022; Weigle, 2010). Among the factors that support successful writing, L2 vocabulary knowledge plays a particularly central role (e.g., Schmitt, 2020). Words are considered to be the basic building blocks of language, required for the construction of larger entities like sentences and paragraphs (e.g., Read, 2000). It is thus no surprise that vocabulary knowledge has emerged as an important predictor of L2 writing performance (e.g., Ellis, 2004; Miralpeix & Muñoz, 2018; Nation, 2013; Webb, 2020).

Vocabulary knowledge is a complex and multifaceted construct (Nation, 2001). L2 learners may be able to understand far more words than they are able to produce in their own speech or writing (Schmitt, 2019). This gap is often captured through the distinction between receptive vocabulary knowledge, the knowledge needed to recognize and understand a word when it is encountered during listening or reading, and productive vocabulary knowledge, the knowledge needed to produce or use a word in speaking or writing (González-Fernández & Schmitt, 2020; Read, 2000; Schmitt, 2010b). Writing thus relies strongly on productive vocabulary knowledge, as it draws on L2 learners' ability to recall word forms from memory based on the meanings they aim to convey (e.g., Bardovi-Harlig, 2020). This makes productive vocabulary knowledge especially crucial to L2 writing.

Nevertheless, writing not only depends on learners' ability to *produce* a word form, but also on their ability to *use* the word in a sentence and broader text context. Vocabulary use in writing provides evidence for how well learners can mobilize their vocabulary knowledge by using it in their own output. In order to investigate vocabulary use in writing, scholars have often taken a corpus-based approach by examining two features of vocabulary in learner texts: lexical diversity (the use of varied words), and lexical sophistication (the use of sophisticated or advanced words) (Bulté

& Housen, 2012; Crossley, 2020; Kyle, 2020a). These constructs have repeatedly been shown to predict L2 writing quality, with human raters awarding higher scores to written texts with a wider range of vocabulary and more low-frequency vocabulary (e.g., Crossley & McNamara, 2012; Daller & Phelan, 2007; Kyle et al., 2018, 2021). At the same time, more proficient learners are known to produce more diverse and more low-frequent vocabulary (e.g., Crossley, 2020; Engber, 1995; Kyle, 2020b; Treffers-Daller et al., 2018). This demonstrates that features of learners' vocabulary use can provide an indication of L2 writing quality and L2 proficiency.

The measurement of lexical sophistication in particular highlights the role of frequency in learners' vocabulary development. Usage-based approaches to language learning have long emphasized that language learning is a result of repeated exposure and use, and word representations in the mental lexicon become stronger as they are encountered more frequently (e.g., Bybee, 1998; Ellis & Collins, 2009; Ellis & Wulff, 2020; Tomasello, 2003). From this perspective, vocabulary development is closely tied to the amount and type of L2 contact learners have. A common source of L2 contact is the L2 classroom, where input is provided in a formal instructional setting. In recent years, research has also increasingly acknowledged the role of out-of-school L2 contact, especially in the context of L2 English (Kusyk et al., 2025; Sundqvist, 2024).

Numerous studies have shown that across educational levels, receptive vocabulary knowledge is supported by both formal instruction (e.g., Peters, Noreillie, et al., 2019; Unsworth et al., 2014; Webb & Chang, 2012) and out-of-school contact (e.g., De Wilde et al., 2020b; Gyllstad et al., 2025; Peters, Noreillie, et al., 2019; Sundqvist, 2019). Some studies have suggested that out-of-school contact may be an even stronger predictor of receptive vocabulary knowledge than formal instruction (e.g., Busby, 2021; Peters, Noreillie, et al., 2019; Warnby, 2022). However, productive vocabulary knowledge and vocabulary use in writing tend to be more demanding because they rely on productive mastery (Schmitt, 2019). Given the limited research into the impact of instruction and out-of-school contact on productive vocabulary knowledge and vocabulary use (for some exceptions, see e.g., Bollansée et al., 2021; Kaatari et al., 2023; Sundqvist & Wikström, 2015), questions remain about their exact relationship.

Furthermore, the amount and type of L2 contact may differ between L2s. Given the unique status of English as a lingua franca and its ubiquity in many European countries (Schmitt, 2019; Sundqvist, 2024; Werner & Tegge, 2021), the results for English cannot be assumed to generalize to other L2s, where the possibilities for out-of-school contact may be less extensive. L2 French, though one of the most commonly taught L2s

(Eurostat, 2024), holds a different status than English, often characterized by a perception of being difficult to learn (Dewaele, 2005a) and less positive attitudes than English (Dewaele, 2005a; Dewaele & Saito, 2024). Flanders, the Dutch-speaking region in Belgium, presents a particularly interesting context in which L2 English and L2 French are both mandatory subjects in secondary school. Formal instruction starts earlier for French (grade 5, ages 10–11) than for English (grade 7 or 8, ages 12–14), but Flemish learners tend to be exposed to much out-of-school English (Peters, Noreillie, et al., 2019) and they also tend to have more positive attitudes towards English (De Clercq & Simoens, 2015; Dewaele, 2005b; Housen et al., 2001). In view of the dominant focus on L2 English in extant research (e.g., Guo et al., 2021; Kusyk et al., 2025; Plonsky, 2023), there is a strong need for cross-linguistic studies to verify whether findings on L2 English also apply to L2 French.

1.1. Rationale and research aims

Given the importance of productive vocabulary knowledge and vocabulary use in L2 writing, this PhD project set out to investigate L2 learners' productive vocabulary knowledge and written vocabulary use in both L2 English and L2 French. While a large body of work has investigated the development of receptive vocabulary knowledge, productive mastery has received less empirical attention, and there is currently only limited understanding of the development of productive vocabulary knowledge (e.g., Schmitt, 2019). In addition, some studies have indicated that the relationship between the development of productive vocabulary knowledge and vocabulary use in writing is complex, and they may not develop in the same way or at the same rate (Henriksen & Danelund, 2015; Laufer, 1998). Yet, little is known about how productive vocabulary knowledge and vocabulary use interrelate and how they develop across grades in secondary school.

Furthermore, the development of vocabulary knowledge and vocabulary use is shaped in part by the learning conditions. Although there is much research on the impact of formal instruction and out-of-school L2 contact on receptive vocabulary knowledge, their relationship with and relative contribution to productive vocabulary knowledge and vocabulary use remain underexplored. This is especially the case in other L2s than English, which has been the focus of the vast majority of previous studies (e.g., Kusyk et al., 2025; Plonsky, 2023; Sundqvist, 2024). In light of the unique status of English as lingua franca and as the language of pop culture (Sundqvist, 2024; Werner & Tegge, 2021), more research is needed to clarify the relationship between L2 contact, productive vocabulary knowledge and vocabulary use in other L2s than English. The situation in Flanders lends itself well to a cross-linguistic approach, as L2 English and L2 French are

both compulsory subjects in secondary education. Yet, L2 learners receive formal instruction for English in a later grade than for French and they are exposed to large amounts of out-of-school English (e.g., Peters, Noreillie, et al., 2019), leading to potentially different learning contexts for these two L2s. As a result, it remains to be seen to what extent findings for L2 English also emerge for L2 French.

In order to address these questions, it is essential to critically examine how productive vocabulary knowledge and vocabulary use are measured, as methodological choices may shape the findings and the conclusions drawn. Since only limited research has taken a multi-dimensional approach to measuring vocabulary use by combining indices of lexical diversity and lexical sophistication (e.g., Consolini & Kyle, 2024), existing studies may not always provide a comprehensive account of learners' vocabulary use. In addition, most studies have adopted only a single approach to examining the use of low-frequency words, based on either frequency counts or frequency bands (see e.g., Crossley et al., 2013). A combined approach may shed more light on the development of learners' vocabulary use in different respects, which is currently especially uncommon in other L2s than English.

Finally, while previous research has investigated the reliability of lexical diversity and lexical sophistication indices (e.g., Crossley et al., 2013; Sung et al., 2024; Treffers-Daller, 2013; Zenker & Kyle, 2021), few studies have paid attention to the pre-processing stage in analyzing learners' vocabulary use. During pre-processing, researchers have to decide what is counted as a single lexical unit (e.g., should *read* and *reading* be counted as occurrences of the same word?) and which words are included in the analyses (e.g., should function words be counted?). While these decisions may have a considerable impact on the findings, few studies have directly compared multiple operationalizations of lexical indices in the pre-processing stage (Maw et al., 2022; Treffers-Daller, 2013; Treffers-Daller et al., 2018).

In light of these research gaps, this PhD project aimed to present a multi-dimensional investigation of L2 learners' productive vocabulary knowledge and vocabulary use in two L2s, English and French. More specifically, this PhD thesis focused on productive vocabulary knowledge measured at the level of form recall (fill-in-the-gap), and vocabulary use as measured through lexical diversity and lexical sophistication in written output. The first objective was to investigate the cross-sectional development of productive vocabulary knowledge and written vocabulary use, as well as their relationship. The second objective was to examine the role of formal instruction and out-of-school L2 contact in productive vocabulary knowledge and written vocabulary use. The third objective was

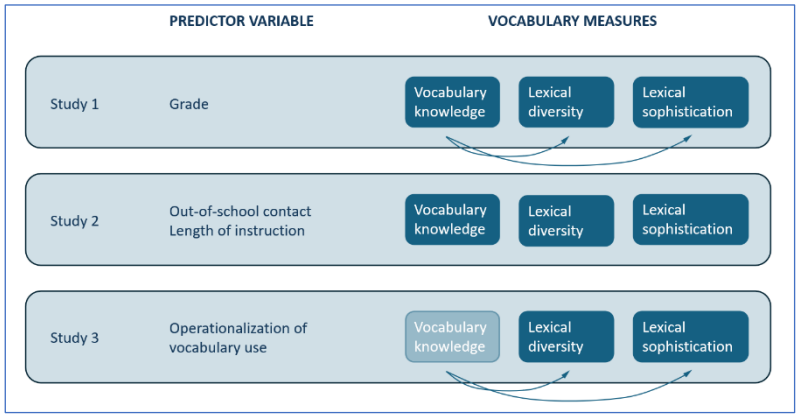
to explore the impact of methodological decisions on findings of vocabulary use. Through three empirical studies (see Figure 1.1.), the PhD project investigated how learners' productive vocabulary knowledge and vocabulary use may be impacted by learners' grade (RQ1), out-of-school L2 contact and length of formal instruction (RQ2), and methodological choices in the operationalization of lexical diversity and lexical sophistication (RQ3). The following general research questions guided the project:

RQ1. How do productive vocabulary knowledge (form recall knowledge) and written vocabulary use (lexical diversity and lexical sophistication) develop cross-sectionally in L2 English and L2 French, and how are they interrelated? (Study 1)

RQ2. What is the role of out-of-school L2 contact and length of formal instruction in productive vocabulary knowledge, lexical diversity and lexical sophistication in written L2 English and L2 French? (Study 2)

RQ3. What is the impact of different operationalizations of lexical diversity and lexical sophistication on the outcomes of lexical analyses in L2 English and L2 French, and how are they related to productive vocabulary knowledge? (Study 3)

Figure 1.1. Overview of the three studies



Data were collected from L2 learners of English and French in grade 10 (ages 15–16 years old) and grade 12 (ages 17–18 years old). Participants were recruited from schools across Flanders (Belgium, L1 = Dutch). We adopted a within-subject design in which each participant completed a test

battery in both English and French. For both L2s, learners completed a productive vocabulary test (productive vocabulary knowledge: form recall knowledge), two informal writing tasks (vocabulary use: lexical diversity and lexical sophistication), and a questionnaire (out-of-school L2 contact, length of formal instruction).

Study 1 investigated the cross-sectional differences between grade 10 and grade 12 learners in productive vocabulary knowledge, lexical diversity and lexical sophistication in written L2 English and L2 French. It statistically compared vocabulary test scores obtained by grade 10 and grade 12 learners, as well as the values for lexical diversity and two indices of lexical sophistication in both grades. Study 1 also focused on the relationship between productive vocabulary knowledge on the one hand, and lexical diversity and sophistication on the other. Through mixed-effects models, we analyzed to what extent productive vocabulary knowledge (and its interaction with grade) predicted the indices of lexical diversity and lexical sophistication.

Study 2 concentrated on the role of out-of-school L2 contact and length of formal instruction in productive vocabulary knowledge, lexical diversity and lexical sophistication in L2 English and L2 French. Through (partial least squares) structural equation modeling, it examined the impact of length of formal instruction and overall out-of-school L2 contact on learners' vocabulary test scores and the indices of lexical diversity and sophistication. In addition, Study 2 also used generalized additive mixed models to examine which types of out-of-school activities were most strongly related to productive vocabulary knowledge and indices of lexical diversity and sophistication.

Study 3 adopted a methodological perspective on the analysis of lexical diversity and sophistication by exploring the effect of four operationalizations related to pre-processing (lemmatization and word inclusion) on the lexical findings in L2 English and L2 French. We ran linear mixed-effects models to compare the lexical values for each operationalization. Study 3 also investigated which operationalization was most strongly related to learners' productive vocabulary knowledge, as a proxy for their lexical proficiency. Generalized additive models were used to investigate the predictive power of productive vocabulary knowledge on each of the operationalizations.

The detailed research questions per Study can be found in Table 1.1.

Table 1.1. Detailed research questions that guided the three studies

Study 1
1. What are the **cross-sectional differences** between grade 10 and grade 12 learners in terms of productive vocabulary knowledge, lexical diversity and lexical sophistication in written L2 English and L2 French? 2. To what extent do productive vocabulary knowledge and grade **predict lexical diversity and lexical sophistication** in written L2 English and L2 French?
Study 2
1. What is the relationship between amount of time spent on **out-of-school contact, length of formal instruction**, gender, and productive vocabulary knowledge and lexical diversity and sophistication in written L2 English and L2 French? 2. Which **types of out-of-school activities** are most strongly related to productive vocabulary knowledge and lexical diversity and sophistication in written L2 English and L2 French?
Study 3
1. To what extent are **lexical findings** for lexical diversity and lexical sophistication in L2 English and L2 French affected by **four different operationalizations** based on lemmatization (default or custom) and word inclusion (all words or content words)? 2. Which operationalization shows the **strongest relationship with learners' productive vocabulary knowledge** as a proxy for lexical proficiency in L2 English and L2 French?

1.2. Structure of the thesis

This PhD thesis is structured as follows. The current chapter (Chapter 1) broadly introduces the research topic, specifies the rationale and research aims of the PhD project, outlines the structure of the thesis, and lists key terminology used in this thesis.

Chapter 2 provides a review of the literature relevant to the three empirical studies included in the thesis. It describes the theoretical and empirical grounds for this PhD project. The literature review is comprised of three parts. The first part focuses on how vocabulary knowledge can be defined and how it can be measured. The second part addresses research that has explored the relationship between vocabulary knowledge and

vocabulary use on the one hand, and L2 skills on the other. Finally, the third part concentrates on how vocabulary can be learned. This third section presents theoretical accounts of vocabulary learning, and then describes research into vocabulary learning in instructional settings as well as through out-of-school L2 contact.

Chapter 3 provides a detailed account of the methodology. The chapter first describes the research design, the research context and the participants. Next, the chapter outlines the selection and development of the data collection instruments, including the pilot study and the final instruments administered to the participants. Finally, the chapter describes the procedure for the data collection and the subsequent lexical analyses.

Chapters 4–6 report on the three empirical studies that were conducted as part of this PhD project. Each study is described in one chapter, which can be read on its own as an independent report of the related literature, lexical and statistical analyses, results and discussion. As a result, there may be some overlap between these chapters and other sections of the thesis. The results of Study 1 (Chapter 4) have been published in *Applied Linguistics*. Studies 2 and 3 (Chapters 5 and 6) have been submitted to international peer-reviewed journals and are currently under review.

Finally, Chapter 7 summarizes the main findings of the three studies. In addition, the chapter highlights the theoretical, pedagogical and methodological contributions of this thesis. The chapter closes with some limitations of the thesis, suggestions for future research, and an overall conclusion.

1.3. Key terminology

Productive vocabulary knowledge

Productive vocabulary knowledge relates to L2 learners' knowledge of a word at a level that allows them to produce the word (Laufer & Nation, 1999; Nation, 2013). This implies they can recall the word form from memory when they are prompted to produce the word (Read, 2000). In this thesis, the notion of productive vocabulary knowledge is used to describe L2 learners' vocabulary knowledge as tested in a form recall test (fill-in-the-gap format, 'controlled' productive vocabulary knowledge). The terms 'productive vocabulary knowledge' and 'form recall knowledge' are used interchangeably to describe the data based on this test format.

Vocabulary use

Vocabulary use is closely related to productive vocabulary knowledge in that it refers to L2 learners' ability to produce a word in the context of their own written or spoken output (Levitzky-Aviad & Laufer, 2013; Nation, 2001). However, vocabulary use refers to a productive context in which learners are not prompted to produce a specific, pre-determined target word. Instead, they can choose freely (not) to use a certain word in their output ('free' productive vocabulary use) (Laufer, 1998; Norris & Ortega, 2001). In this thesis, vocabulary use was analyzed by means of two constructs: lexical diversity and lexical sophistication.

Lexical diversity

Lexical diversity is a measure of vocabulary use that is related to the use of varied words (Carroll, 1938; Kyle, 2020b; Read, 2000). It is typically operationalized as the proportion of types (unique, unrepeated words) to tokens (all words) in a text, which represents the degree of (lexical) repetition in a text. Since the traditional type-token ratio has been criticized for its dependence on text length, the present thesis used the moving-average type-token ratio (MATTR) as an index of lexical diversity (Covington & McFall, 2010; Zenker & Kyle, 2021).

Lexical sophistication

Lexical sophistication is a measure of vocabulary use that is related to the use of sophisticated words. Lexical sophistication has traditionally been defined as the use of low-frequency words, because frequency is an important predictor of vocabulary knowledge and L2 proficiency (Kyle, 2020b; Read, 2000). While lexical sophistication is now more commonly regarded as a multi-dimensional construct (Kim et al., 2018a; Nicklin et al., 2025), the present thesis adopts a frequency-based interpretation of lexical sophistication. It provides a nuanced analysis of frequency-based lexical sophistication by employing two indices (Crossley et al., 2013): a band-based measure representing the proportion of low-frequency words in a text (lexical frequency profile), and a mean-based measure representing the average frequency of the words in a text (average log frequency).

Out-of-school L2 contact

Out-of-school L2 contact refers to any contact with an L2 that is initiated by the L2 learner, outside of or independent from the L2 classroom (Arndt, 2019; Sundqvist, 2009). Although other terms have been coined to refer to similar notions, in the current thesis, I will use the term 'out-of-school contact' because of three main reasons: (a) it does not imply any activity or passivity on the part of the learner, unlike, for instance, 'out-of-school exposure'; (b) it does not refer to one specific L2, unlike, for instance,

'extramural English'; and (c) it does not inherently assume that learning takes place during the activities, unlike, for instance, 'informal second language learning'.

Length of formal instruction

Length of formal instruction relates to the total amount of formal instruction L2 learners have received (Milton, 2009; Miralpeix, 2020), expressed in terms of the number of years. Formal instruction, in this thesis, is restricted to the school environment. It refers only to L2 classes in school and does not include learner-initiated activities outside of school, even if these include a formal or instructional approach to teaching an L2 (e.g., classes during a language camp, Duolingo).

Lemmatization

Lemmatization is a text pre-processing technique that is applied prior to lexical analyses (e.g., lexical diversity or lexical sophistication) being conducted (Maw et al., 2022; Treffers-Daller et al., 2018). Lemmatization typically means that inflected word forms (e.g., *see, sees, seeing, saw, seen*) are clustered under one base form, the lemma (e.g., *see*). All occurrences of the inflected word forms are then counted as occurrences of the same lemma (Kyle, 2020b; Nation, 2013; Webb, 2021). In Study 3, we also explored a custom lemmatization scheme that distinguished verbal lemmas based on mood and tense, thus reducing the number of word forms under one lemma (see Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024).

Word inclusion

Word inclusion is another decision in the pre-processing stage (Engber, 1995; Kyle & Eguchi, 2023). It is used in this thesis specifically to refer to the inclusion or exclusion of function words in lexical analyses. In Study 3, we compared lexical findings for the inclusion of all words, including function words (i.e., words carrying grammatical meaning, like determiners, prepositions), compared to the inclusion of content words only (i.e., words carrying lexical meaning). In Study 3, 'all words' is used to refer to analyses that have been performed on both function words and content words, and 'content words' is used to refer to analyses of content words only (Kyle et al., 2018).

1.4. Bibliographical information

Study 1 (Chapter 4): Vocabulary knowledge and vocabulary use in writing: A cross-sectional comparison of L2 English and L2 French.

Published in *Applied Linguistics*.

Reference:

Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., Puimège, E., Creten, S., & Peters, E. (2025). Vocabulary knowledge and vocabulary use in writing: A cross-sectional comparison of L2 English and L2 French. *Applied Linguistics, Advance online publication,* amaf009. https://doi.org/10.1093/applin/amaf009

Study 2 (Chapter 5): The role of out-of-school language contact and instruction in L2 English and L2 French vocabulary knowledge and use.

Under review.

Reference:

Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., & Peters, E. (under review). The role of out-of-school language contact and instruction in L2 English and L2 French vocabulary knowledge and use.

Study 3 (Chapter 6): Measuring vocabulary use in L2 English and L2 French writing: How methodological decisions shape the results.

Revised and resubmitted.

Reference:

Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., & Peters, E. (revised and resubmitted). Measuring vocabulary use in L2 English and L2 French writing: How methodological decisions shape the results.

2. Background literature

This chapter provides an overview of the background literature relevant to the three empirical studies included in this PhD thesis. There may be some overlap between the literature described in this chapter and in the literature reviews of the empirical studies, because the studies can be read independently. The current chapter is divided into three parts. The first part (Vocabulary knowledge) discusses the construct of vocabulary knowledge, including how it can be defined and how it can be measured. The second part (Vocabulary knowledge and skills) outlines research on the relationship between vocabulary knowledge and L2 skills (listening, reading, speaking, writing). The third part (Vocabulary learning) homes in on how vocabulary is learned. It first discusses theoretical underpinnings of vocabulary learning. Secondly, it discusses existing research on vocabulary learning in instructional settings and in out-of-school contexts.

PART I: Vocabulary knowledge

2.1. Defining vocabulary knowledge

This section focuses on how the construct of vocabulary knowledge can be defined. The first subsection deals with what a word is and the ways in which words can be counted. The second subsection describes what it means to know a word. The third subsection discusses different dimensions or degrees of mastery.

2.1.1. What is a word?

A key question in L2 vocabulary research is what is considered a 'word'. In Richards and Schmidt (2002), a word is defined as 'the smallest of the linguistic units which can occur on its own in speech or writing' (p. 588). Specifically in writing, words are 'strings of letters separated by white space and/or punctuation' (Kyle, 2020a, p. 333). The question is how broadly the notions of a 'linguistic unit' or 'string of letters' in these definitions should be interpreted. For example, should contracted forms (e.g., *don't, I've, c'est*) be counted as one or more words (*don't* or *do not*)? Should the contracted form be counted as a distinct form from the full form (*n't* or *not*, *c'* or *ce*)? If learners are able to produce the words *livre* (English: *book*) and *livres* (English: *books*) on a test, can they be said to know two words or two forms of the same word?

When counting the words in a text, a distinction can be made between *tokens* and *types* (e.g., Kyle, 2020b; Read, 2000). The number of tokens is the total number of all running words that appear in a text. For instance, the previous sentence consists of seventeen tokens. The number

of types, on the other hand, is the number of different or unique word forms in a text. A type is each word that has not appeared in the text before. For instance, the sentence *The boy saw the other boy* contains six tokens, but only four types (*the, boy, saw, other*), because the second occurrences of *the* and *boy* are repetitions of tokens that have already appeared. The number of types in a text can be strongly affected by which word forms are counted as occurrences of the same headword.

The way in which words are counted is essential to research on vocabulary knowledge, because it shapes how vocabulary knowledge is estimated and tested, how research findings are interpreted, and how vocabulary goals are formulated. There are four common counting units that can be distinguished in L2 vocabulary research: the orthographic form, the lemma, the flemma, and the word family. In the simplest form, the counting unit can be the orthographic form, in which case the lexical unit under inspection is the exact word form as it appears in a text or on a test. Each orthographic form is then counted as a separate type (e.g., related forms like *act, acts, acted, acting, react, reacts, active, actively* are counted as separate headwords). Using the orthographic form as a counting unit implies that learners are assumed to have no knowledge at all of the inflectional or derivational system of an L2. The orthographic form may therefore underestimate many L2 learners' knowledge unless they are at a true beginner level (Kyle, 2020b).

Two broader counting units are the lemma and the flemma, which are essentially word clusters comprised of the stem or the base form of a word (the headword) and its inflections, but not derivations. For example, the verb forms *act*, *acts, acting, acted* are all grouped together under the headword *act*, but derivations like *react, active, inactive, actively* do not fall under this same headword. The lemma distinguishes between parts of speech, which means that the verb *act* and the noun *act* are categorized as two separate headwords, each with their own inflections. The flemma is slightly more comprehensive than the lemma because it does not distinguish between parts of speech, that is, *act* (verb) and *act* (noun) with their respective inflections are categorized under the same headword. In practice, the difference between flemmas and lemmas may be small because only a limited number of headwords have the same orthographic form in multiple parts of speech (Jarvis & Hashimoto, 2021).

Finally, word families are word clusters comprised of the stem of a word, its inflections and its derivations. For example, the word forms *act, react, reacts, reacted, active, inactive, actively* are all grouped together under the headword *act*. This counting unit relies on the assumption that L2 learners can recognize morphological structures and that they can see that words from the same word family are related (e.g., Webb, 2021b). Using the

word family as the counting unit implies that learners who know one form in the word family, also know all of other the inflected and derived forms (Kremmel, 2016). While it may be reasonable to expect that learners are able to recognize or understand different derivational forms when they encounter them (e.g., Snoder & Laufer, 2022, though see McLean, 2018), it may be less realistic to assume they can produce or use all of these forms (Nation & Webb, 2011; Schmitt, 2010a; Webb, 2021b). L2 learners often have partial receptive knowledge of derivations and affixes (Kremmel & Schmitt, 2016; McLean, 2018; Sasao & Webb, 2017), but they tend to have only limited productive knowledge of all the derivational forms of a headword (Iwaizumi & Webb, 2022; Schmitt & Zimmerman, 2002). For instance, it cannot be assumed that if learners can produce the form *central*, they can also produce the word *decentralization*, or by extension, use the word in their own output. Lemmas and flemmas may thus be more suitable counting units for L2 production than word families because they assume less extensive knowledge of the morphological and derivational system (e.g., Gablasova & Brezina, 2021; Kyle, 2020b).

The choice of the counting unit can have an effect on the number of types that are distinguished in a text. Table 2.1. illustrates the different counting units based on a sentence with five occurrences of the headword *run* in English (the base form *run*, inflected forms like *running* and derivative forms like *runner*), and five occurrences of the headword *marche* in French (the base form *marche*, inflected forms like *marchent* and derivative forms like *marcheur*). Both example sentences contain five tokens based on their respective headwords. However, based on the comparison of the different counting units, it becomes apparent that the choice of the counting unit affects the final number of types.

Table 2.1. Illustration of the different counting units for the headwords *run* (English) and *marche* (French)

During his run, the runner was running faster than all the other runners had run before. (5 tokens based on the headword *run*)
Orthographic form: 4 types 1. *run* (N) (V) 2. *runner* 3. *running* 4. *runners*
Lemma: 3 types 1. *run*: run (N) 2. *run*: run (V), running 3. *runner*: runner, runners

Flemma: 2 types
1. *run*: run (N), running, run (V)
2. *runner*: runner, runners
Word family: 1 type
1. *run*: run (N), runner, running, runners, run (V)
Pendant la marche, le marcheur marche plus vite que les autres marcheurs ne marchent. (5 tokens based on the headword *marche*)
Orthographic form: 4 types
1. *marche* (N) (V)
2. *marcheur*
3. *marcheurs*
4. *marchent*
Lemma: 3 types
1. *marche*: marche (N)
2. *marcher*: marche (V), marchent
3. *marcheur*: marcheur, marcheurs
Flemma: 2 types
1. *marche*: marche (N), marche (V), marchent
2. *marcheur*: marcheur, marcheurs
Word family: 1 type
1. *marche*: marche (N), marcheur, marche (V), marcheurs, marchent

Note. (N) = noun; (V) = verb

2.1.2. Knowing a word

Scholars have long tried to define what it means to 'know' a word. What seems to be widely agreed upon, is that vocabulary knowledge is multifaceted and complex in nature (e.g., Nation, 2013; Read, 2000; Richards, 1976; Schmitt, 2010a). Nation's (2001, 2013, 2022) framework is probably one of the most cited frameworks to describe the components of vocabulary knowledge. Nation distinguishes three aspects of vocabulary knowledge: *form*, *meaning* and *use*. Each of these is comprised of receptive and productive dimensions of knowledge[1]. Receptive knowledge is related to the ability to understand or recognize a word, and generally associated with the skills of listening and reading, whereas productive knowledge is related to the ability to produce a word, typically associated with the skills of speaking and writing (e.g., Nation, 2013; Schmitt, 2010b).

In terms of form, receptive knowledge entails recognizing what the word sounds like (pronunciation) and looks like (spelling), as well as which

[1] Some scholars have also referred to receptive and productive vocabulary as 'passive' and 'active' vocabulary knowledge respectively (e.g., Fan, 2000; Laufer, 1998; Meara, 1990).

parts the word is comprised of (e.g., morphological features). Productive knowledge of form means being able to produce the word with the correct pronunciation, accurate spelling and the appropriate word parts (e.g., the plural noun for *kid* is *kid<u>s</u>*, but the plural noun for *child* is *child<u>ren</u>*; the plural noun for *enfant* is *enfant<u>s</u>* [children], but the plural of *jeu* is *jeu<u>x</u>* [games]).

In terms of meaning, Nation defines receptive knowledge as knowing what meaning the word signals, to which concept it refers, and which words are associated with the word (e.g., the word *child* is related to words like *parent, family, sibling, birth, care, cry, school*). Productive knowledge of meaning entails being able to recall a word form for a certain concept (e.g., the word *child* to refer to a young human in a stroller), being able to produce the word in a range of contexts (e.g., to describe children of different ages), and being able to produce synonyms (e.g., *kid, toddler*) and antonyms (e.g., *adult*).

The third component of vocabulary knowledge is use. This component is related to how words are used in context, that is, in a sentence or in a text. Receptive knowledge of use means being able to recognize that a word has been used grammatically correctly in a sentence, that it tends to co-occur with certain other words (e.g., collocations), and in which contexts (e.g., register, connotation) or how often (e.g., frequency) the word is likely to appear. Part of knowing a word is, thus, to assess the likelihood of a word appearing in a specific sentence construction or, more broadly, in a text (Richards, 1976). Productive knowledge of use is related to the ability to use the word in spoken or written output in correct grammatical patterns (e.g., transitivity of a verb), in combination with words that commonly occur with the word (e.g., multi-word units), and in an appropriate context (e.g., degree of formality). Productive knowledge of use then means that L2 learners are able to make an informed decision (yet not necessarily a conscious decision) on whether or not to use a word in their own output. This is henceforth referred to as vocabulary use.

Godfroid (2020) has proposed expanding Nation's (2013) framework by adding a new dimension, namely, automaticity. Automaticity underpins all three components of vocabulary knowledge (form, meaning, use) and it reflects how fluently learners can access their knowledge of the aforementioned vocabulary components. This echoes Laufer and Nation (2001), who also proposed taking into account fluency as a dimension of vocabulary knowledge. While this is in line with an increasing interest in process-oriented and 'online' measures of vocabulary knowledge in the field, the present thesis adopts Nation's original framework because it aligns well with the focus of the empirical studies.

2.1.3. Dimensions of mastery

Nation's (2013) framework distinguishes two dimensions of knowledge for each of the components: receptive and productive knowledge. Webb (2008) demonstrated that the better learners' receptive vocabulary knowledge, the better their productive vocabulary knowledge. For highly frequent words, words that were known receptively were often also known (partially) productively. However, there seems to be a consensus that productive vocabulary knowledge of form, meaning and use represents a level of mastery that is more difficult to reach, and therefore more advanced, than receptive vocabulary knowledge (Nation, 2013). This is supported by studies that have shown that learners' receptive vocabulary knowledge is larger than their productive vocabulary knowledge (e.g., Fan, 2000; Laufer, 1998; Noreillie et al., 2020; Webb, 2008). In addition, Webb's (2008) study showed that for less frequent words, learners could not always demonstrate productive knowledge for words they knew receptively.

The development from receptive to productive vocabulary knowledge and vocabulary use is not straightforward. In a study on different components of vocabulary knowledge, Laufer (1998) compared receptive vocabulary knowledge (multiple-choice test), productive vocabulary knowledge (fill-in-the-gap test) and free vocabulary use (writing task) of L2 learners in grades 10 and 11 (16- and 17-year-olds). Her study showed that there was a larger gap between receptive and productive vocabulary knowledge in grade 11 than 10, indicating that productive knowledge developed at a slower pace than receptive knowledge. The study also found that learners with more low-frequency vocabulary knowledge did not necessarily use more low-frequency vocabulary in their writing output. These findings point to a consistent 'order of acquisition', such that when knowledge is viewed as a continuum from receptive to productive mastery, receptive aspects of vocabulary knowledge tend to develop before productive aspects (e.g., González-Fernández & Schmitt, 2020; Melka, 1997; Miralpeix, 2020).

It is important to note that there is a lack of a widely agreed upon conceptualization of receptive and productive vocabulary knowledge (e.g., Melka, 1997; Read, 2000, 2020; Webb, 2005). As a result, researchers have operationalized receptive and productive vocabulary knowledge in many different ways. For instance, tests requiring learners to explain or define an L2 target word in their L1 requires them to retrieve the meaning from memory. While this seems to represent relatively 'more productive' knowledge than recognizing the correct definition among several given response options, it does not require the learner to produce anything in the L2. Based on Nation and Beglar (2007, p. 11), who state that receptive

knowledge entails being able to access the meaning of a provided word form, it could be argued that such a test is a test of receptive rather than productive knowledge, and productive knowledge is measured through tests that require the learner to produce an L2 word.

To avoid any confusion about terminology, researchers have turned to a different distinction to describe dimensions of mastery, namely, the distinction between recognition and recall. Recognition is related to learners' ability to recognize the form or meaning of a word, whereas recall is directed at learners' ability to recall a word form or meaning from memory (e.g., Read, 2020). Laufer and Goldstein (2004) tested L2 learners' knowledge of the form-meaning link of the exact same words at four different levels of mastery (form recognition, meaning recognition, form recall, meaning recall). They found that the four vocabulary test formats yielded different test scores. Learners performed best on the form recognition test and the meaning recognition test, followed by the meaning recall test and finally the form recall test. The test scores showed moderate to large correlations (r = 0.31–0.65, p <.001), which suggests that the test formats measured related but distinct constructs, in line with other research on the construct of vocabulary knowledge (e.g., González-Fernández & Schmitt, 2020; Stewart et al., 2024). Using implicational scaling, González-Fernández and Schmitt (2020) demonstrated that recognition knowledge was consistently acquired before recall knowledge, both for the form-meaning link as well as for other components of vocabulary knowledge (derivations, homonymy/polysemy, collocations). This was confirmed in a comparative study of L2 learners of English with an L1 Chinese and an L1 Spanish background (González-Fernández, 2025). These studies support the idea that vocabulary development is incremental, and vocabulary knowledge at the level of recognition tends to precede knowledge at the level of recall. Although Webb (2013) notes that this is not always the case (e.g., a single exposure to a word may allow learners to produce the correct word form, even if they are not able to recognize the meaning), it is conceivable that recall knowledge requires more extensive and more active knowledge than recognition knowledge (e.g., Nation, 2013).

Finally, a distinction can be made between vocabulary breadth and vocabulary depth (Nation, 2013; Qian & Lin, 2020; Read, 2000; Schmitt, 2014; Webb, 2013, 2020). Vocabulary breadth (or vocabulary size) refers to the range or number of words an L2 learner knows at a certain level of mastery (e.g., meaning recognition, form recall). Vocabulary depth, on the other hand, is related to the quality of that knowledge, or how well an L2 learner knows those words (Anderson & Freebody, 1981; González-Fernández, 2021; Schmitt, 2014; Yanagisawa & Webb, 2020). For instance, one could measure vocabulary depth by testing knowledge of the form-

meaning link, collocations, associations and grammatical functions[2] of the same set of words (e.g., Schmitt, 2014; Webb, 2013). Qian and Lin (2020, p. 68) compare vocabulary depth to the process of advancing knowledge of a word on a continuum from one end, superficial familiarity with a word, to the other end of the continuum, correct use in free production. Vocabulary breadth and depth are closely related, but they are best regarded as separate constructs and they can be measured in different ways (Koizumi & In'nami, 2020; Yanagisawa & Webb, 2020). Given the focus of the present PhD thesis, vocabulary depth will not be further discussed in detail.

In sum, Section 2.1. introduced the concept of *words* and how they can be interpreted and counted (e.g., types/tokens, counting units). The section also illustrated that vocabulary knowledge is a multifaceted construct, encompassing various aspects (e.g., form, meaning, use) and dimensions (e.g., receptive and productive knowledge, recognition and recall, vocabulary breadth and depth).

2.2. Measuring vocabulary knowledge

This section will focus on the different ways in which vocabulary knowledge can be measured. It discusses test formats commonly used to measure vocabulary breadth and introduces several well-known and widely used vocabulary tests.

2.2.1. Vocabulary test formats

Tests of vocabulary breadth have commonly focused on measuring knowledge of the form-meaning link (Schmitt, 2014). This section describes some common test formats measuring learners' knowledge of the form-meaning link, although this by no means denies the existence of test formats probing into other aspects of vocabulary knowledge (e.g., derivational or collocational knowledge).

It is important to note that the test format is dependent on the type of knowledge the test is envisioned to measure. A vocabulary test that requires test takers to select a correct definition from four options, may not reveal anything about the test takers' ability to provide their own definition or explanation, nor about their ability to produce the form based on the definition alone. Studies such as those by Laufer and Goldstein (2004) and González-Fernández and Schmitt (2020) have demonstrated that learners'

[2] Webb (2005) illustrates that tests of grammatical functions are essentially tests of part-of-speech knowledge. In his receptive test of grammatical functions, learners had to select one out of three sentences that contained a target word in the correct part-of-speech (e.g., used as a noun, verb or adjective). The productive test of grammatical functions required learners to produce their own sentence using the target word in the correct part-of-speech.

knowledge of a word at one level of mastery does not imply their knowledge at other levels of mastery.

There are four common test formats to measure vocabulary breadth, often operationalized as knowledge of the form-meaning link. A distinction can be made between recognition and recall tests, also referred to as selected-response and constructed-response tests (e.g., Read, 2020). Both recognition and recall can be measured at the level of form (recognizing or producing a word form) or meaning (recognizing or describing a meaning) (Laufer & Goldstein, 2004).

2.2.2. Selected-response vocabulary tests

Form recognition

The first type of selected-response test is a form recognition test. Form recognition tests measure learners' knowledge of a word form. A common test format is a checklist of target words, whereby test takers have to tick a checkbox for each word for which they believe to know the meaning (e.g., Meara & Jones, 1988). Form recognition tests can also take a multiple-choice format, in which case test takers are expected to select the correct word form from multiple options with different orthographic forms (e.g., Pellicer-Sánchez, 2016; Webb, 2005).

An example of an English vocabulary test at the level of form recognition is the Eurocentres Vocabulary Size Test (Meara & Jones, 1988), also referred to as the Yes/No Test or the checklist format. It requires learners to self-assess and indicate whether they know the meaning of a word, usually by checking the boxes of the words they believe to know (see below). An important advantage of this test format is that it is very quick and easy to complete. However, the format has also been criticized for relying on self-report data by learners who have to gauge their own knowledge without actually demonstrating that knowledge (e.g., Read, 2000, 2020). The test tries to account for this by including pseudowords: for each frequency band, both real and non-existent words are shown. The proportion of non-existent words that learners have checked, is accounted for mathematically in the final test score. Still, the test is considered to likely represent an overestimation of learners' vocabulary knowledge. Moreover, the binary distinction in the Yes/No Test does not allow much nuance regarding the extent of learners' (self-reported) knowledge of the target items.

Example of a Yes/No Test batch (reduced for illustration purposes):

☐ galpin ☐ impulse ☐ suggest ☐ advance

☐ peculiar	☐ benevolate	☐ indicate	☐ needle
☐ destruction	☐ compose	☐ ager	☐ debt
☐ generate	☐ fast	☐ buttle	☐ horobin
☐ route	☐ undertake	☐ descript	☐ attach

Meaning recognition

The second type of selected-response test is a meaning recognition test, commonly measured in a multiple-choice format. Meaning recognition is probably the degree of mastery most commonly investigated, and most validated vocabulary tests are tests of meaning recognition (e.g., Stewart et al., 2024). Multiple-choice tests of meaning recognition require test takers to select the correct meaning of a provided target word from several definitions (Nation & Beglar, 2007; Peters, Velghe, & Van Rompaey, 2019). This type of test can also be offered in a matching format, which provides test takers with clusters of target words and definitions (e.g., Batista & Horst, 2016; Nation, 1983; Schmitt et al., 2001). Every cluster is comprised of more target words than definitions, and test takers have to match the correct target words with their respective definitions.

Two well-known tests of vocabulary breadth at the level of meaning recognition in English are the Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001) and the Vocabulary Size Test (Nation & Beglar, 2007). The Vocabulary Levels Test (VLT) was first developed by Nation in the 1980s and was further improved and validated by Schmitt et al. (2001). It presents a clustered matching format, requiring test takers to link a set of three definitions to three out of six possible L2 words (for an example, see below). The VLT is a frequency-based test containing items sampled from frequency bands with the 2,000, 3,000, 5,000 and 10,000 most frequent word families, as well as a section on academic vocabulary. Each section contains 6 clusters, totaling 30 target items per section (and 150 items in total).

Example of a VLT cluster:

1. accident	
2. debt	_____ loud deep sound
3. fortune	_____ something you must pay
4. pride	_____ having a high opinion of yourself
5. roar	
6. thread	

More recently, a new and improved version of the VLT was published: the updated Vocabulary Levels Test (Webb et al., 2017). The updated VLT addresses some limitations of the previous version of the VLT. A first adaptation is related to the frequency levels of the items, which were drawn from the 1,000, 2,000, 3,000, 4,000 and 5,000 most frequent word families. The 1,000 most frequent word families were deemed important because they account for a large portion of written and spoken English (approximately 65 to 85%, Webb et al., 2017). The 4,000 frequency level was also included because this resulted in a vocabulary test of the 5,000 most frequent word families. As 'these are the most important words for learners' (Webb et al., 2017, p. 36), the combination of these five frequency levels was deemed more informative to estimate learners' vocabulary knowledge than the inclusion of the 10,000 level and academic vocabulary. A second adaptation involved an update of the frequency lists from which the items were sampled, which were selected to better reflect current English (the British National Corpus/Corpus of Contemporary American English word list was used). A final adaptation can be found in the presentation of the test to increase transparency to the test takers (see below).

Example of an updated VLT cluster:

	capital	career	committee	exam	fence	option
choice						
job						
test						

Another widely used vocabulary test of meaning recognition in English is the Vocabulary Size Test (VST) (Nation & Beglar, 2007). An important difference between the VLT and VST is that the VST provides an indication of learners' total vocabulary size (at the level of meaning recognition), rather than testing their knowledge of particular frequency levels. The VST provides test takers with a target word form in a non-defining sentence and four definitions (see below), only one of which provides a correct description of the target word meaning. The other three options are distractors that 'do not share core elements of meaning with the correct answer' (Nation, 2012, p. 4). The VST tests knowledge of words among the 14,000 most frequent word families, with 10 target items per 1,000-word frequency band. The items were sampled from the British National Corpus based on Bauer and Nation's (1993) word list up to Level 6. More recently, two parallel 100-item versions were created, based on the 20,000 most frequent word families (Nation, 2012). The estimate of learners' total vocabulary size is calculated by multiplying their test score by 100 (140-item test) or by 200 (100-item test).

Example of a VST item:

soldier: He is a **soldier**.
a. person in a business
b. student
c. person who uses metal
d. person in the army

There is also a picture-based vocabulary size test, the Picture Vocabulary Size Test (Anthony & Nation, 2021), which can be used with beginner and low-literacy learners (e.g., primary school children). The items were derived from a corpus comprised of books and magazines for children, movies and TV shows, and spoken American and British English. The Picture Vocabulary Size Test contains 96 items from the 6,000 most frequent word families (although there is also a version up to 20,000 word families which can be used with pre-literate native speakers).

The meaning recognition tests discussed above are all aimed at measuring vocabulary knowledge in English, and there are few validated tests available for other L2s, like French. One French vocabulary test at the level of meaning recognition is the *Test de la Taille du Vocabulaire* (TTV; Batista & Horst, 2016), which is a meaning recognition test modeled after the VLT (Nation, 1983; Schmitt et al., 2001; see below). It tests learners' knowledge of words at the 2,000, 3,000, 5,000 and 10,000 lemma-based frequency levels. The items up until the 5,000 frequency level were available from Lonsdale and Le Bras (2009), while the items at the 10,000 level had to be sampled from an older list by Baudot (1992). The TTV does not include a section on academic vocabulary like the VLT, because no list of academic vocabulary existed for French at the time of the development of the TTV. In addition to the TTV, there is also a picture-based test of French vocabulary knowledge at the level of meaning recognition, the *Échelle du Vocabulaire en Images de Peabody* (Dunn et al., 1993). This is a 239-item picture-based test meant for young learners, similar to the aforementioned PVST (Anthony & Nation, 2021).

Example of a TTV cluster:

1. concours
2. division _____ grand plaisir
3. joie _____ un moyen de transport
4. phase _____ séparation en deux parties
5. stade
6. véhicule

The final tests of meaning recognition that should be mentioned are the VocabLab tests, developed by Peters, Velghe, and Van Rompaey (2019) in parallel versions for both L2 English and L2 French, with a focus on learners with an L1 Dutch background. The VocabLab tests are frequency-based meaning recognition tests in the same format as the VST (Nation & Beglar, 2007). The items were sampled from the lemma-based British National Corpus/Corpus of Contemporary American English frequency list for the English test (Davies, 2008), and Lonsdale and Le Bras' (2009) frequency list for the French test. The tests are comprised of a total of 120 items, equally divided across the 2,000, 3,000, 4,000 and 5,000 most frequent lemmas. The 2,000 frequency level contains items from the first and second 1,000 most frequent lemmas. The test developers created both monolingual (completely in English or French) and bilingual (response options in Dutch) versions (see below for an example of each). The aim of the VocabLab tests was to address several critiques of the VLT and VST: (a) the risk of guessing was reduced through an 'I don't know' option; (b) the items were sampled from more recent frequency lists; (c) cognates with L1 Dutch were excluded to avoid overestimation of vocabulary knowledge.

Example of a French meaning recognition VocabLab test item:

Monolingual test:	*Bilingual test:*
quitter	quitter
☐ être entraîné vers le sol	☐ vallen
☐ avoir droit à quelque chose	☐ verdienen
☐ laisser quelqu'un en partant	☐ verlaten
☐ se composer de plusieurs parties	☐ bestaan uit
☐ je ne connais pas la réponse	☐ ik weet het niet

2.2.3. Constructed-response vocabulary tests

Meaning recall

In addition to selected-response tests, there are also constructed-response tests, which require test takers to produce their own response rather than to select a given response from several options. The first type of constructed-response test is a meaning recall test. A straightforward way of testing meaning recall is to provide test takers with an L2 word form and asking them to describe the meaning (Read, 2020). The meaning can be expressed either in the L2, through an L2 definition or L2 synonym, or in the L1, through an L1 description or L1 translation of the target word (e.g., González-Fernández & Schmitt, 2020; Webb, 2005). There seem to be few widely used and validated meaning recall tests, although the Computer-

Adaptive Test of Size and Strength (Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004) contains a section where learners are asked to provide an L1 translation or definition of the L2 target item.

Example of a CATSS meaning recall test item:

Laufer & Goldstein (2004):
affluence _____

Aviad-Levitzky et al. (2019):
They have a lot of time.
This word means: _____

Form recall

The second type of constructed-response test is a form recall test. In a form recall test, learners have to recall a target word by retrieving the word form from memory. A common form recall test format is the fill-in-the-gap format, which usually presents test takers with isolated sentences containing a missing word (e.g., Laufer & Nation, 1999; Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019). Based on the sentence context, test takers are expected to recall the word form of the missing word. Another form recall format is a translation test in which test takers have to produce the L2 word form based on L1 definitions or translations of the intended target word (e.g., Kuppens, 2010; Webb, 2005, 2008). A more demanding variation of a form recall test is one where test takers are asked to produce a full sentence using the target word form, although this format is less commonly used because it is more difficult to score (Read, 2020).

The best-known test of vocabulary breadth at the level of form recall is probably the Productive Levels Test (Laufer & Nation, 1999). As the name suggests, it is inspired by the VLT, and consists of 18 items at each of the frequency levels (2,000, 3,000, 5,000, 10,000 most frequent word families and academic vocabulary). The Productive Levels Test (PLT) is a fill-in-the-gap test which provides test takers with a meaningful sentence context with one missing word (see below). A meaningful sentence context may help the learner retrieve the correct word form because it more closely resembles conditions in which the L2 form may normally be encountered (Miralpeix, 2020; Nation, 2013). The first letters of the missing word are provided to eliminate the possibility of learners responding with another word that may suit the context, but that is from a different frequency level. The number of letters was decided for each item separately, based on how many letters were necessary to avoid the possibility of any alternative

responses being elicited. The length of the line at the end of the missing word does not serve as an indication of the number of missing letters.

Example of a PLT item:

The book covers a series of isolated epis_____ from history.

More recently, the aforementioned VocabLab tests were also developed in a form recall test format to measure learners' vocabulary knowledge in L2 English and French at the level of form recall (Jamers, 2021; Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019). The same target items were used in the form recall tests as in the meaning recognition tests, which means that the tests are also lemma-based. This is likely a more appropriate counting unit than the word family to test productive knowledge, especially for learners that may not be advanced.

The productive VocabLab tests take a similar format to the PLT (Laufer & Nation, 1999), albeit with a few notable differences (see below for a sample item). First, the productive VocabLab tests present an L1 (Dutch) translation of the envisioned target word at the end of the sentence[3]. This reduces the number of potential responses, and it decreases the impact of receptive vocabulary knowledge or reading comprehension on learners' ability to recall the correct word form (see Waring, 2002, as cited in Nation, 2013). This is in line with Nation's (2013) recommendation to use bilingual tests of vocabulary knowledge, especially for lower proficiency learners (Elgort, 2013; Nguyen & Nation, 2011). It is also supported by some studies that have shown that L2 learners were able to demonstrate better vocabulary knowledge on bilingual than monolingual vocabulary tests (De Clercq, 2016; Elgort, 2013; Vrancx, 2016). Second, the VocabLab tests use underscores to indicate the exact number of letters missing from the target words, which also lowers the possibility of alternative responses. Third, the missing words in the sentences are displayed with as few letters as possible. The number of letters required is often lower than in the PLT (which sometimes displays half of the target word; e.g., *car*__ for *career*, *popu*__ for *popular*, *oa*__ for *oath*), because the range of possible alternative responses is reduced by the L1 translation and the indication of the number of remaining letters. In practice, this means that for most items, only one or two letters have to be provided.

[3] The bilingual version (with L1 translations) of the productive VocabLab test in French was already available (see Noreillie, 2019). The productive VocabLab test in English, however, did not yet include L1 translations. These were added for this PhD project. More information on this procedure can be found in Chapter 3.

Example of a French form recall VocabLab test item:

Le pôle Sud, c'est l'e_ _ _ _ _ le plus froid de la planète. *(plaats)*

Finally, it should be mentioned that in addition to tests of vocabulary breadth as described above, there are also tests of vocabulary depth. Well-known examples are the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996), the Word Associates Test (Read, 1993; Schoonen & Verhallen, 2008), and Lex30 (Meara & Fitzpatrick, 2000). These will not be discussed in detail here as they are beyond the scope of the PhD thesis.

To summarize, Section 2.2. provided an overview of common test formats to measure vocabulary breadth (form and meaning recognition, form and meaning recall), and introduced some well-known tests that have been used in vocabulary research (e.g., EVST, VLT, PLT, TTV, VocabLab). This section demonstrated that many widely used vocabulary tests measure meaning recognition knowledge, and the vast majority of available tests are aimed at L2 learners of English.

2.3. Measuring vocabulary use

The previous section focused on the measurement of vocabulary knowledge through discrete-point tests, where learners demonstrate knowledge of specific pre-determined target items. Another aspect of vocabulary knowledge is *use* (Nation, 2013): the ability to use vocabulary in speech or writing. The subsections below will first describe common indices to measure L2 learners' vocabulary use, after which some considerations will be addressed with regard to how the learner texts are (pre-)processed. Finally, the role of topic and instructions will be briefly acknowledged.

2.3.1. Indices of lexical richness

Although discrete-point tests of vocabulary knowledge provide valuable information about learners' knowledge of the form-meaning link, they are limited in scope in that they do not measure how well learners can use the words in their own spoken or written output (Nation, 2013). Yet, an important aspect of mastering vocabulary is the ability to effectively use it in communicative situations (Read, 2000; Schmitt, 2014; Webb, 2020). Vocabulary use is therefore often measured through spoken or written texts generated by L2 learners based on a communicative prompt. In the classification of test formats as selected-response or constructed-response, this may be considered a format of 'free constructed response' (Norris &

Ortega, 2001, p. 169), or 'free active vocabulary' (Laufer, 1998, p. 257)[4]. Such a collection of learner texts (i.e., the learner corpus) provides a foundation for subsequent corpus-based analyses. Usually, some form of computational modeling is then used to analyze vocabulary use in the learner corpus (Ellis & Wulff, 2020). Some measures of vocabulary use require a comparison of the vocabulary in the learner corpus to that in a so-called reference corpus: a large-scale collection of texts representing authentic speech or writing, usually comprised of L1 texts by native speakers (e.g., Corpus of Contemporary American English, Davies, 2008; though see Monteiro et al., 2020 for a call to use L2 rather than L1 benchmarks).

Vocabulary use is most commonly measured as lexical richness, typically divided into three subconstructs related to word choice: lexical density, diversity and sophistication (e.g., Bulté & Housen, 2012; Crossley, 2020; Durrant et al., 2022; Kyle, 2020b; Read, 2000). These constructs are also included in Bulté and Housen's (2012) model of lexical complexity, which distinguishes systemic lexical complexity (density and diversity) from structural lexical complexity[5] (sophistication). The sections below describe how lexical density, diversity and sophistication can be measured.

Lexical density

The first subconstruct is lexical density, which is related to the information density in a text. It is most commonly operationalized as the relative proportion of lexical or content words in a text (i.e., open-class words that carry lexical meaning, like nouns, lexical verbs, adjectives and adverbs), as opposed to function words (i.e., closed-class words that carry grammatical meaning, like prepositions, pronouns and determiners). A set of two example sentences is shown below, both consisting of thirteen tokens. In sentence 1, there are six lexical words (underlined), whereas sentence 2 contains eight lexical words. This means that sentence 2 has a larger proportion of lexical words and is thus considered to have higher lexical density.

[4] This test format can be contrasted with 'constrained constructed response' or 'controlled active/productive vocabulary', which requires test takers to produce a short response, often a pre-determined target word form or meaning (e.g., in form and meaning recall tests).

[5] Bulté and Housen (2012) also include compositionality as an aspect of structural lexical complexity, which refers to the morphemes and syllables words are comprised of. This subconstruct will not be further elaborated because it (a) refers to sub-word level components of vocabulary use, which are not directly relevant to the present thesis focusing on vocabulary use of single words, (b) is inconsistent with common working definitions of a closely related construct, lexical richness.

Lexical density

1. The dog chased the cat, but the cat was faster than the dog.
→ 6 lexical words / 13 total words (tokens) = 0.46

2. The canine chased the feline, but the agile cat outpaced the pursuing hound.
→ 8 lexical words / 13 total words (tokens) = 0.62

Measures of lexical density seem to be useful to distinguish between spoken and written modes (Ure, 1971), more and less interactive spoken tasks (O'Loughlin, 1995), and text genres (Biber, 2012). Lexical density tends to be lower in spoken, interactive and informal texts. However, while lexical density has emerged as a reliable method to distinguish between text genres and between spoken or written modalities, there is a lack of clear empirical evidence of the relationship between lexical density and L2 writing performance (Durrant et al., 2022; Kyle, 2020b). While some studies have tried to link lexical density to speaking skills (Hyltenstam, 1988; Lu, 2012) and writing skills (Engber, 1995; Hyltenstam, 1988; Linnarud, 1986), few studies seem to have yielded a significant relationship between them (for an exception, see De Clercq, 2015), indicating that lexical density may not be the most appropriate measure of L2 skills (Consolini & Kyle, 2024; Kyle, 2020b).

Moreover, Durrant et al. (2022) posit that lexical density is essentially not a measure of vocabulary use. To illustrate this, they compare two constructions like *big black woolly sheep* and *sheep that are big and black and have a lot of wool*. These sentence clauses use largely the same lexical words (*big, black, woolly/wool, sheep*), but the sentences are formulated with a different number of function words, thus affecting lexical density. Durrant et al. (2022, p. 64) conclude that lexical density represents 'the different syntactical ways in which information is packaged' and as such, it is a matter of syntax rather than a matter of vocabulary use. For these reasons, lexical density will not be further discussed in this PhD thesis.

Lexical diversity

The second subconstruct of lexical richness is lexical diversity, which is related to the variety of words in a text. The concept of lexical diversity was first introduced by George Kingsley Zipf in 1935, when he claimed humans use language to express concepts in a way that strives for an optimal balance between variety and repetition (Jarvis, 2017). The specific term *diversity*, however, was first coined by Carroll (1938) when he

introduced his concept of *diversity of vocabulary*. At a minimum, the term refers to the level of lexical repetition in a text, interpreted as the 'relative amount of repetitiveness' or the 'relative variety in vocabulary' (Carroll, 1938, p. 379). Some scholars have related this to the range of words in a text and the concept of vocabulary breadth (Bulté & Housen, 2012; Olinghouse & Wilson, 2013). Nonetheless, Jarvis (2013, 2017) has noted that lexical diversity should be interpreted more broadly than lexical repetition, namely, as a hypernymic construct comprised of seven properties: volume (number of unique words, or types), abundance (total number of words, or tokens), variety (relative proportion of unique words), dispersion (intervals between repeated words), evenness (degree of equality in the number of repetitions across types), specialness (use of words that enhance perceived diversity, often related to the level of usefulness of information), and disparity (semantic and orthographic similarity of words). Since reliable measures of constructs broader than lexical repetition are still under development (Jarvis, 2017; Kyle et al., 2021), this section will focus on the existing indices to measure lexical repetition, which align with Jarvis' concepts of volume, abundance, and variety[6].

Lexical diversity has been operationalized in many ways, but measures of lexical diversity are intended to reflect the range of different, unrepeated words that learners use in their output. The most traditional and perhaps most well-known measure of lexical diversity is the type-token ratio (Johnson, 1944). The type-token ratio (TTR) is a value between 0 and 1 that represents the proportion of unique, unrepeated words (types) to the total number of words (tokens) in a text (see the example sentences below). The TTR value starts at 1 and decreases as the number of repetitions increases (see Table 2.2.). The original TTR has been strongly criticized for its dependence on text length. Shorter texts tend to have higher TTR values because they require less repetition (e.g., fewer function words need to be repeated). In longer texts, however, 'the number of unused types available to a writer decreases as they continue to write' (Durrant et al., 2022, p. 65). This can be problematic, since more proficient learners are known to produce longer texts on average (e.g., Bulté & Housen, 2014; Crossley, Kyle et al., 2015; Treffers-Daller, 2013). Their proficiency may, however, not be reflected in the measurement of lexical diversity through TTR, because they are penalized for writing longer texts.

Lexical diversity: original type-token ratio

1. The dog chased the cat, but the cat was faster than the dog.

[6] Some of the indices discussed also touch on Jarvis' concept of dispersion, albeit indirectly. For instance, as will be explained later, the window length of the moving-average type-token ratio sets a limit on the distributional properties of words in a text by constraining repetitions to a pre-determined window length.

→ 8 types / 13 tokens TTR = 0.62

2. The canine chased the feline, but the agile cat outpaced the pursuing hound.
→ 10 types / 13 tokens TTR = 0.76

Table 2.2. Illustration of decreasing TTR with increased repetitions

Yesterday, **our** dog chased **a** cat into **the** barn.	TTR = 1.00
Yesterday, **our** dog chased **the** cat into **the** barn.	TTR = 0.89
Yesterday, **the** dog chased **the** cat into **the** barn.	TTR = 0.78

One way to resolve the effect of text length on research findings based on TTR is to analyze only part of the learner output (e.g., 200 words from each learner text). However, since this implies a loss of data and potentially valuable information, scholars have more commonly tried to resolve the dependency on text length mathematically (Daller & Xue, 2007; Durrant et al., 2022; Kyle, 2020b), for instance, through Guiraud's Index (Root TTR), Corrected TTR, Maas' Index (Log TTR), D, and HD-D.

Guiraud's Index (Root TTR; Guiraud, 1960) is a simple variation on the original TTR, in that it takes the square root of the denominator (i.e., the number of tokens). Corrected TTR (Carroll, 1964) is calculated in a similar way, but it takes the square root of double the number of tokens. Maas' Index (Log TTR; Maas, 1972) accounts for text length differences by fitting the value to a logarithmic curve: it is calculated using the log-transformed numbers of tokens and types[7]. These adapted TTR measures have been found to be related to L2 writing proficiency (e.g., Bulté & Housen, 2014; Durrant et al., 2022; Treffers-Daller et al., 2018), although they still seem to be dependent on text length (Daller & Xue, 2007; Malvern et al., 2004; Zenker & Kyle, 2021). D and HD-D are two other measures that have been developed in an attempt to address the text length dependency issue more robustly. D (also referred to as vocd-D; Malvern et al., 2004) is a complex measure that calculates the TTR for random text segments of 35–50 tokens of increasing size, and generates an empirical curve of the relationship between TTR and window length (the number of tokens in the segment). The shape of this curve is fit against a theoretical curve by the coefficient *D*. However, McCarthy and Jarvis (2007) found that D was essentially an approximation of the probability of words occurring in a sample from the text, and it is still dependent on text length. HD-D

[7] These indices are calculated with the following formulas. TTR: ntypes / ntokens. Guiraud's Index: ntypes / √ntokens. Corrected TTR: types / √(ntokens*2). Maas' Index: (log(ntokens) – log(ntypes)) / log(ntokens)^2. Note that for Maas' Index, lower values indicate higher diversity.

(McCarthy & Jarvis, 2010) was therefore developed as an improved version of D. It uses the hypergeometric distribution to calculate the probability of a word in the text occurring in random samples of 42 tokens. HD-D, thus, represents the sum of probabilities of each word in a text appearing in a random sample at least once.

In recent years, some studies have tried to determine the reliability of lexical diversity indices by investigating their dependence on text length (e.g., Kyle et al., 2021, 2024; McCarthy & Jarvis, 2010; Sung et al., 2024; Treffers-Daller, 2013; Zenker & Kyle, 2021). Two indices have emerged as particularly stable measures of lexical diversity across text lengths: the measure of textual lexical diversity (MTLD) and the moving-average type-token ratio (MATTR). MTLD (McCarthy & Jarvis, 2010) is calculated by determining the mean number of sequential words at which a pre-determined TTR is maintained, before it falls below that value. The mean number of words is calculated based on sequential, non-overlapping text segments. The pre-determined TTR is usually set to 0.72 because this is the value at which TTR values tend to stabilize (McCarthy & Jarvis, 2010). MTLD is thus represented by a natural number representing the mean number of words it takes before the TTR falls below 0.72. Research has shown that MTLD is relatively independent of text length and it is stable even with short English texts of less than 100 words (e.g., Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010). This has been confirmed in a study on L2 Korean writing (Sung et al., 2024) and L2 French speaking (Treffers-Daller, 2013), although the latter study also showed that D and HD-D outperformed MTLD in predicting general L2 proficiency.

A final measure of lexical diversity is the moving-average type-token ratio (MATTR). MATTR (Covington & McFall, 2010) is computed as the mean TTR based on the TTR values of a moving window in a text. More concretely, the TTR is calculated for contiguous, overlapping, same-size text segments until all words in a text have been included in the calculation. For instance, with a window size of 50 words (common for the analysis of written L2 texts), the TTR would be computed for text segments of words 1–50, then for words 2–51, for words 3–52, for words 4–53, and so forth, until the last word of the text has been included[8]. The average TTR based on all of those values returned, is the final MATTR value. The chosen window size should be sufficiently large to provide a realistic picture of repetition in the analyzed texts. For example, a window size of 2 would not be very informative, because the likelihood of 2 sequential words being the

[8] There is also another variant of this measure, the mean segment type-token ratio (MSTTR). MSTTR is computed very similarly to MATTR, except that the TTR values are calculated for non-overlapping text segments of a certain window size (e.g., words 1–50, 51–100, 101–150 etc.). Once the number of remaining words in a text is smaller than the window size (e.g., 40 words left), these words are discarded.

same is rather low (so MATTR would likely be nearly 1 for all texts). On the other hand, the window should also be small enough to allow TTR values to be calculated for multiple text segments. If, for example, the analyzed texts are only around 100 words, it would not be possible or logical to use a 200-word window size. In studies on L2 vocabulary use, researchers have often opted for a window size of 50 words for written texts, and a window size of 11 words for spoken texts because smaller window sizes seem to be more strongly related to oral proficiency ratings (Kyle et al., 2024). This means that the shortest possible L2 text that can be analyzed, should consist of at least 50 (written text) or 11 (spoken text) tokens.

MATTR has been shown to be particularly stable across text lengths (Kyle et al., 2024; Sung et al., 2024; Zenker & Kyle, 2021). Two example sentences are shown below, where MATTR is calculated based on 5-word segments. It is useful to compare these values to the aforementioned values for the original TTR. A comparison clearly shows that the sentences appear to differ much more in terms of lexical diversity based on TTR (0.62 and 0.76) than based on MATTR (0.89 and 0.91). The actual difference is, however, a small one: sentence 1 contains only 2 more repetitions than sentence 2. In short texts, a few more repetitions can make a large difference for TTR, which can distort the interpretation of the results. This effect seems to be somewhat neutralized with MATTR, which still shows higher diversity for sentence 2 than sentence 1, but it more accurately reflects the actual difference in lexical repetition.

Comparative studies have consistently ranked MATTR among the top-performing indices of lexical diversity. In addition to its stability across text lengths, a study by Jarvis (2017) also showed that MATTR was more strongly associated with human judgements of diversity than MTLD (MATTR $r = 0.58$; MTLD $r = 0.48$). Several scholars have thus recommended using MATTR as an index of lexical diversity (e.g., Bulté et al., 2025; Zenker & Kyle, 2021). However, in spite of its reliability, not many studies to date have operationalized lexical diversity through MATTR (Kyle et al., 2024; though with some exceptions, e.g., Kyle et al., 2021; Kyle & Eguchi, 2023; Woods et al., 2023). Research using MATTR is also limited in other L2s than English (see Díez-Ortega & Kyle, 2024; Tracy-Ventura et al., 2021, for Spanish; Sung et al., 2024, 2025, for Korean), and in particular in L2 French.

Lexical diversity: moving-average type-token ratio

1. The dog chased the cat, but the cat was faster than the dog.
   → MATTR = 0.89

2. The canine chased the feline, but the agile cat outpaced the pursuing hound.
   → MATTR = 0.91

Lexical sophistication

The third subconstruct of lexical richness is lexical sophistication. Lexical sophistication is not related to the range of different words in a text, but rather to the nature of words in a text. It refers to the use of sophisticated or advanced words (or, by extension, multi-word units). The level of 'sophistication' of a word depends on multiple word-related properties that can affect how difficult a word is to learn (e.g., Ellis, 2002; Kyle, 2020b). Word properties associated with sophistication include (but are not limited to) frequency, psycholinguistic word information (e.g., concreteness, familiarity, imageability, age of acquisition), contextual distinctiveness (i.e., the variety of contexts in which a word can be used), and *n*-gram strength of association (consecutive *n*-grams or dependency *n*-grams[9]) (e.g., Kim et al., 2018; Kyle, 2020b). These measures have all been shown to be related to L2 writing proficiency and holistic writing quality scores. Especially word frequency, bigram frequency and bigram strength of association have featured as strong predictors in multiple studies (e.g., Consolini & Kyle, 2024; Kim et al., 2018; Kyle et al., 2018).

However, the traditional and prevalent interpretation of lexical sophistication is frequency-based, and lexical sophistication is still commonly operationalized in line with this traditional frequency-based interpretation (e.g., Bulté & Housen, 2012; Kyle, 2020b; Laufer & Nation, 1995). Frequency is strongly correlated with other word-related properties, like concreteness, imageability and age of acquisition (Divjak & Caldwell-Harris, 2015). These other word properties may also contribute to the sophistication of a word, and thus, how difficult or advanced it may be (see also Ellis, 2002). Nonetheless, there is clear evidence of frequency as one of the main predictors of L2 vocabulary knowledge and L2 proficiency (e.g., Boone & De Wilde, 2023; Consolini & Kyle, 2024; Kyle et al., 2018; Nation, 2006; Puimège & Peters, 2019a). In addition, frequency predicts word processing speed during lexical decision tasks (e.g., Brysbaert & New, 2009; Chen et al., 2018; Whaley, 1978), response times on meaning recognition tests (e.g., Laufer & Nation, 2001), and eye-tracking measures of word processing speed during reading (e.g., Kuperman et al., 2024; Nahatame & Uchida, 2025). These findings further support the notion that frequency is an essential word-related variable in predicting L2 proficiency.

There are two general ways in which researchers have previously operationalized the use of low-frequency words: a band-based approach

[9] *N*-grams refer to *n* number of words co-occurring together. Consecutive *n*-grams are words that co-occur chronologically, that is, they appear next to each other (for bigrams, e.g., *they offered me a job*). Dependency *n*-grams are words that co-occur syntactically, that is, they have a syntactic relationship (for bigrams, e.g., *they offered me a job*). The more likely words are to co-occur, the higher their strength of association.

and a mean-based approach. First, the band-based approach has its roots in a study by Laufer and Nation (1995), which introduced the Lexical Frequency Profile (henceforth LFP). The LFP is a measure of frequency-based sophistication that draws on frequency bands in word lists compiled from a reference corpus. To calculate the LFP, each word in a learner text is checked for its ranking in a reference frequency list (e.g., among the first 1,000 most frequent words, or '1K'; among the 1,001–2,000 most frequent words, or '2K'). In Laufer and Nation's study, the frequency list consisted of three bands: 1K, 2K, and academic vocabulary (University Word List). Words not appearing in any of these bands, were classified as off-list. Laufer and Nation's study showed that the LFP could discriminate between proficiency levels and that it was stable across two texts by the same learners, providing support for the validity and reliability of the measure.

Because the LFP provides an overview of the proportion of words at each frequency level, it can be difficult to interpret, and researchers have sometimes used a 'condensed LFP' (e.g., Laufer, 1995, 1998). The condensed LFP is based on the same calculations, but it returns a single value representing the proportion of words beyond a certain frequency level (e.g., beyond 2K), which is interpreted as the cut-off point between 'basic and beyond-basic vocabulary' (Laufer, 1995, p. 267). In Laufer and Nation (1995) and Laufer (1998), this meant the sum of the proportion of words in the University Word List and off-list words, meaning that words at the 1K and 2K levels were regarded as high-frequency and basic, and words beyond 2K were considered low-frequency and sophisticated. Other research using the LFP has often adhered to this original interpretation and has typically reported results beyond 2K or up to the 2K level (e.g., Henriksen & Danelund, 2015; Johnson et al., 2016; Laufer, 1998). Moreover, learners who demonstrate meaning recognition knowledge of the 2K level tend to perform adequately on listening, reading, and writing tasks (Stæhr, 2008), supporting the interpretation of words up to 2K as basic vocabulary.

The LFP has been criticized for being coarse-grained because it is based on 1,000-word bands (e.g., Crossley et al., 2013; Meara, 2005a). This means that if a learner uses a word ranked as the 1,001 most frequent word, this is interpreted as equal to the use of a word ranked as the 1,999 most frequent word (both 2K). Meara (2005a) concludes from his Monte Carlo Analysis that the LFP may be appropriate to distinguish between learners with highly different vocabulary sizes, but it is not sensitive enough to detect smaller changes in vocabulary size. In a response to Meara's critical analysis, Laufer (2005b) clarifies that the LFP cannot be interpreted as a measure of vocabulary *size*, because it is a measure of vocabulary *use*. The two are not to be confused: learners who have gained knowledge of more low-frequency words, may not necessarily (be able to) use these words in their own free production (see also Laufer, 1998).

In recent years, researchers have increasingly used a mean-based approach to analyzing frequency-based lexical sophistication in English (e.g., Crossley et al., 2013; Kyle & Crossley, 2015). The mean-based approach is characterized by the use of word frequency counts based on a reference corpus, without ranking the words in order of decreasing frequency and dividing them into 1,000-word bands. Instead, the frequency counts from the reference corpus are used directly to compute the average frequency of the learner texts. These corpus frequency counts are often normalized per million words to account for differences in corpus size between corpora, or to make the numbers 'more manageable' (Crossley et al., 2013, p. 967). The average frequency can be logarithmically transformed to account for skewness that is typical of frequency counts (Zipf's law).

The mean-based approach has been suggested to be more fine-grained than the band-based approach, because it does not draw on 1,000-word bands (Kyle, 2020b). In fact, Crossley et al. (2013) compared lexical findings based on the LFP to those based on average frequency. They demonstrated that the average frequency was more accurate in classifying texts according to their proficiency level (58%) than the LFP, although it was followed by the LFP with an accuracy rate of 48%. Yet, few studies have combined mean-based and band-based indices of lexical sophistication, and more research is needed to deepen our understanding of these indices.

2.3.2. Pre-processing learner texts

The previous section provided an overview of the ways in which research has commonly measured lexical diversity and lexical sophistication in L2 learner texts. It showed through different indices *how* learners' vocabulary use can be measured, but it did not discuss *what* exactly is measured. This section will discuss two methodological decisions researchers have to make in order to compute the aforementioned indices: how words in a text are lemmatized, and which words are included in the analyses.

Lemmatization

Lemmatization is a text pre-processing technique that clusters occurrences of inflected word forms under one headword, the lemma (see also Section 2.1.1.). Researchers nowadays are becoming increasingly aware of the impact of the level of morphological richness in an L2 on lemmatization, although limited research to date has accounted for this in lexical analyses (e.g., Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024). In languages with a rich inflectional system, such as Spanish or French, much more extensive knowledge of word forms is assumed on the part of the learner than for less morphologically rich languages. For instance, the

verbal lemma *run* in English comprises four forms (*run*, *runs*, *running*, *ran*), while the verbal lemma *courir* in French comprises over twenty forms (*courir*, *cours*, *court*, *courons*, *courez*, *courent*, *courais*, *courait*, *courions*, *couriez*, *couraient*, *courrai*, *courras*, *courra*, *courrons*, *courrez*, *courront*, *courrais*, *courrait*, *courrions*, *courriez*, *courraient*, *courant*, *couru*, among others).

Few studies to date have attempted to account for morphological richness in their lexical analyses. An exception is Consolini and Kyle (2024), who adopted a lemmatization scheme for L2 Spanish that did not categorize all inflected verb forms under one headword, but rather under different headwords for each tense and mood combination (e.g., a separate lemma for the indicative present and the subjunctive present tense, but also for the indicative present and the indicative past tense). As a result, these customized lemmas did not consist of as many word forms as the standard lemma would. Their findings based on the custom lemmatization were in line with research on L2 English, with more proficient learners producing more varied and less frequent vocabulary. However, they did not directly apply a similar scheme to L2 English, nor did they compare the results of the custom lemmatization to a standard lemmatization scheme. Since little research has accounted for the level of morphological richness in the L2, more empirical evidence is necessary to determine the impact of the operationalization of lexical items on research findings.

Word inclusion

Another consideration in the analysis of vocabulary use is related to the words which are included in the calculations. Specifically, a distinction can be made between content words (i.e., lexical words) and function words (i.e., grammatical words) (Kyle, 2020b). Content words are open-class words that carry lexical meaning. They are usually considered to include nouns, lexical verbs, adjectives, and adverbs (Engber, 1995; Kyle, 2020b; Lu, 2012; Reid, 1990). Function words are words with a mainly grammatical function, such as determiners, pronouns, prepositions and conjunctions. Copular and auxiliary verbs (e.g., *be*, *have*, *do*) are also typically interpreted as function words (Kyle, 2020b). Function words are known to be highly frequent and they make up a large part of the 1,000 most frequent words in English (Öztürk, 2022; Vilkaitė-Lozdienė & Schmitt, 2020). They also account for large proportion of words in texts: Nation (2013) states that about half of the words in texts tend to be from the 176 most frequent word families, and they are virtually all function words. Function words are required 'in almost every sentence to fulfil the grammatical requirements of the language' (Nation, 2013, p. 231).

As a result, function words are also frequently repeated, leading to a higher token count (more words in total) but a proportionally lower type

count (more repetitions, so fewer unique words). It stands to reason that these characteristics of function words may affect research findings in terms of lexical diversity and frequency-based lexical sophistication. Indeed, analyses based on content words or function words have produced disparate results (Durrant et al., 2019), and content word frequency has been shown to predict (to some extent) ratings of L2 writing tasks (e.g., Crossley et al., 2011; Kyle et al., 2018). However, little is still known about how the inclusion or exclusion of function words in lexical analyses may affect findings.

2.3.3. The role of the topic and instructions

It is important to note that while the aforementioned measures of lexical diversity and sophistication seem to be reliable indicators of L2 writing performance and L2 proficiency, learners' vocabulary use can be shaped to a certain extent by the prompt and the instructions provided. For instance, an early study by Reid (1990) showed that L2 English learners produced a different proportion of content words on average depending on the topic. More recently, Yoon (2017) demonstrated that topics that were relevant to the learners' own experiences, that is, they were more familiar to the learners, elicited more complex vocabulary. The topic had a large-sized effect on learners' use of low-frequency words, measured through the average frequency. In a similar vein, Zenker and Kyle (2021) compared lexical diversity indices across two topics and found that one topic elicited more diverse vocabulary than the other, as measured by means of MATTR. The difference was small, but statistically significant. These topic effects have been widely established in L2 writing research (e.g., Schoonen, 2005, 2012). This seems to imply that in order to reduce potential topic effects on findings of lexical analyses, it is important to administer multiple writing tasks to learners when measuring vocabulary use. This is currently not common practice in L2 vocabulary research (for some exceptions, see Sung et al., 2025; Zenker & Kyle, 2021).

Furthermore, the instructions that accompany a writing task can also impact learners' writing output. At a minimum, clarity of instructions is a basic condition for a valid test (Schoonen, 2011). In order to make sure that the instructions are well understood, Schoonen (2011) proposes that instructions can be given in the learners' L1 if the research sample or student population allows it (i.e., they share the same L1). This can be useful to ensure that the test is actually measuring what it is intended to measure (e.g., a writing task should measure learners' writing performance and not depend on their L2 reading comprehension to understand the instructions). This suggestion aligns closely with Nation's (2013) recommendation on the use of bilingual tests of vocabulary knowledge. In

addition, instructions should clearly inform the test takers about what is expected of them so they have a clear writing goal and target audience in mind (Graham & Perin, 2007; Schoonen, 2011). For a writing task, this could mean that test takers are informed about content properties of the task (e.g., the topic they are expected to write about; the addressee) and about practical restraints (e.g., the time limit; the minimum or maximum length of the text; whether the use of resources, such as dictionaries, is allowed, prohibited or encouraged).

Another consideration is that the instructions given to the learners may also contribute to the extent to which learners' performance on a writing task reflects their underlying competence (R. Ellis, 2004). In line with Cronbach's (1970) distinction between *ability* and *typical behavior*, output on a writing task may not always represent learners' true ability (see also Schoonen, 2011). While it remains difficult to determine whether choices in vocabulary use are a result of underlying competence/ability or rather stylistic choices (Crossley, 2020), the task design and instructions can increase the likelihood of learners' writing output reflecting their competence or ability. For instance, selecting topics that are engaging and interesting to the learners is likely to support their motivation and sustained effort (Van Steendam et al., 2025). In addition, shaping the instructions so that they encourage learners to perform to the best of their abilities may help to strengthen learners' motivation and reduce their anxiety, factors which have been shown to affect L2 performance (e.g., Al-Hoorie, 2018; Dewaele et al., 2023; Dörnyei, 2005; Tahmouresi & Papi, 2021).

To sum up, Section 2.3. discussed the various considerations that need to be made when measuring L2 vocabulary use. While a wide array of indices are available to measure lexical diversity, few studies have adopted indices that are robust against variation in text length (e.g., MATTR), especially in other L2s than English. The section also addressed the value of combining a band-based and mean-based approach to measure lexical sophistication, although little research so far has adopted such a combined approach. Further, the section discussed two considerations in the pre-processing stage that have received limited empirical attention (lemmatization and word inclusion), even though they could strongly affect lexical findings regarding diversity and sophistication. Finally, the section addressed the potential impact of the writing task and the importance of administering multiple writing tasks to measure L2 writing performance.

PART II: Vocabulary and skills

2.4. Vocabulary knowledge and skills

As the focus of this PhD thesis is on vocabulary knowledge and vocabulary use in writing, this section provides an overview of research into the relationship between L2 vocabulary knowledge and the four language skills. The receptive skills are listening and reading, and the productive skills are speaking and writing (e.g., Clenton & Booth, 2022; Qian & Lin, 2020). The first subsection focuses on the relationship between vocabulary knowledge and the receptive skills (reading and listening), while the second subsection focuses on the relationship between vocabulary knowledge and the productive skills (speaking, writing).

2.4.1. Reading and listening

Research has investigated the relationship between L2 vocabulary knowledge and the four L2 skills as a proxy for language proficiency. Most studies investigating the relationship between vocabulary knowledge and the skills have measured receptive vocabulary knowledge, most often using meaning recognition tests. Receptive vocabulary knowledge has consistently been linked to reading comprehension (e.g., Anderson & Freebody, 1981; Miralpeix & Muñoz, 2018; Qian, 1999; Qian & Lin, 2020; Schmitt et al., 2011; Stæhr, 2008). Stæhr's (2008) study showed a strong correlation between VLT scores of low-level L2 English learners and reading comprehension (r = 0.83), which was a stronger correlation than for listening comprehension. In a similar vein, Miralpeix and Muñoz (2018) tested form recognition knowledge of upper-intermediate to advanced learners of L2 English, measured through X_Lex and Y_Lex (Meara, 2005b; Meara & Miralpeix, 2006). They also established a strong correlation between vocabulary test scores and reading comprehension, with form recognition knowledge explaining 27% of the variance in reading test scores. In an investigation of the relationship between reading comprehension and vocabulary knowledge at the level of meaning recognition and meaning recall, Laufer and Aviad-Levitzky (2017) showed that scores on both vocabulary tests were strongly related to reading comprehension scores. These studies all indicate that vocabulary knowledge is indispensable for successful L2 reading.

Apart from reading comprehension, receptive vocabulary knowledge has also been shown to predict performance on L2 listening tests (e.g., Han & Qian, 2024; Miralpeix & Muñoz, 2018; Noreillie et al., 2018; Stæhr, 2008, 2009; Wallace, 2022). Using the VLT as a measure of vocabulary knowledge, Staehr (2009) found that vocabulary knowledge predicted 50% of the variance in listening test scores of advanced L2

English learners. In a replication study with intermediate learners of L2 English and L2 French, Noreillie et al. (2018) confirmed this positive relationship between vocabulary knowledge (measured by the receptive VocabLab tests) and listening comprehension, showing that vocabulary knowledge explained 40% of the variance in the English listening scores and 49% of the variance in the French listening scores. A strong relationship between receptive English vocabulary knowledge and listening comprehension ($r = 0.63$) has also been found for lower-level learners (e.g., Staehr, 2008). Moreover, a study by Vafaee and Suzuki (2020) found that English vocabulary knowledge (measured as breadth using an aural version of the VST, and as depth using a word association test) was almost twice as strong a predictor of listening comprehension than syntactic knowledge ($\beta = 0.55$ compared to $\beta = 0.28$), underscoring the crucial role of vocabulary knowledge in L2 listening skills.

Given the importance of vocabulary knowledge for reading and listening comprehension, scholars have attempted to estimate how much vocabulary knowledge is necessary for comprehension. This topic has been extensively studied in lexical coverage research, at least for L2 English. Lexical coverage is the proportion of words in a given text that are known by the learner (e.g., Webb, 2021a). Research has indicated that L2 learners need to know approximately 95% of running words in a text to achieve basic reading comprehension, whereas 98% coverage has been recommended for comfortable unassisted reading (Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011). For listening, research has indicated that a slightly lower coverage of at least 90% may suffice for 'adequate' comprehension, while 98% coverage better supports deeper understanding (van Zeeland & Schmitt, 2013). Research by Durbahn et al. (2020) has suggested that coverage levels for watching audiovisual content tend to be in line with those for listening, although a more recent study suggested that 80% coverage may suffice for 'minimal adequate' viewing comprehension (Durbahn et al., 2024).

In terms of vocabulary knowledge required to reach an optimal coverage level of 98%, Nation (2006) posits that L2 learners of English should know approximately 8,000 to 9,000 word families for reading comprehension and 6,000 to 7,000 word families for listening comprehension. For 95% coverage, an objective of 4,000 to 5,000 known word families has been suggested for reading comprehension (Qian & Lin, 2020). For listening comprehension, Van Zeeland and Schmitt (2013) have proposed that 2,000 to 3,000 word families may suffice, and 3,000 word families may also suffice for viewing comprehension (Webb & Rodgers, 2009). However, it should be noted that recent research has questioned the generalizability of these figures, since the number of word families required for comprehension may depend on the difficulty and the genre of the text

in question (e.g., Kremmel et al., 2023; Noreillie et al., 2018), and lexical coverage was recently found to explain only limited variation in reading comprehension scores (Webb et al., 2025).

2.4.2. Speaking and writing

Given the focus of the present thesis, it is important to discuss research into the role of vocabulary knowledge in productive L2 skills, that is, speaking and writing. Similar to listening and reading, receptive vocabulary knowledge has emerged as an important predictor of speaking and writing skills (e.g., Lin, 2023; Miralpeix & Muñoz, 2018; Noreillie et al., 2020; Qian & Lin, 2020; Stæhr, 2008; Yang et al., 2019). For instance, a study with L2 learners of French found that receptive vocabulary knowledge at the level of meaning recognition was moderately correlated with L2 speaking performance (Noreillie et al., 2020). A similar relationship has been established between form recognition knowledge in L2 English and speaking performance (Uchihara & Clenton, 2020). In terms of L2 writing, studies have indicated that receptive vocabulary knowledge at the level of meaning recognition was a strong predictor of writing performance in L2 English (e.g., Lin, 2023; Stæhr, 2008). Miralpeix and Muñoz (2018) demonstrated that writing was the L2 skill best predicted by form recognition knowledge of advanced L2 English learners. These findings clearly illustrate the importance of receptive vocabulary knowledge for productive skills.

Furthermore, although the research on productive vocabulary knowledge is more limited, some research has also looked into the role of productive vocabulary knowledge and the L2 skills of speaking and writing (e.g., Johnson et al., 2016; Karakoç & Köse, 2017; Noreillie et al., 2020; Uchihara et al., 2020). Some studies have indicated that productive vocabulary knowledge at the form recall level significantly predicts L2 speaking performance by beginner to intermediate learners of L2 English (Koizumi & In'nami, 2013) and by intermediate to advanced learners of L2 Dutch (de Jong et al., 2012). However, a study with low-intermediate learners of L2 French found no significant correlations between form recall knowledge and speaking performance (Noreillie et al., 2020). The authors explain the lack of a relationship in light of the low proficiency of the L2 learners, which might have caused limited variation in the data for the form recall test. Studies using other measures of productive vocabulary knowledge, like the word association test Lex30 (Meara & Fitzpatrick, 2000), have produced mixed findings: Uchihara et al. (2020) established a positive relationship between productive vocabulary knowledge and global speaking skills (rated as the level of comprehensibility), while Uchihara and

Saito (2019) found no correlations between vocabulary knowledge and global speaking skills.

Only few studies have looked at the relationship between productive vocabulary knowledge and writing. Although not all of these studies have observed a significant relationship (e.g., Yang et al., 2019), most have established (significant) moderate correlations between L2 writing performance and productive vocabulary knowledge in L2 English, as measured through form recall tests (Johnson et al., 2016; Karakoç & Köse, 2017) and word association tests (Lin, 2023). The study by Karakoç and Köse (2017) measured productive vocabulary knowledge of Turkish learners of L2 English through a reduced version of the PLT (Laufer & Nation, 1999) that only tested the 2,000 frequency level. The results indicated a moderate correlation (r = 0.43) between productive vocabulary knowledge and L2 writing performance, and vocabulary test scores explained 18% of the variance in writing scores. These results suggest that tests of relatively high-frequency words can still predict writing performance fairly well, although Qian and Lin (2020) also note that the 3,000 most frequent word families may play a more limited role in more advanced learners' writing skills. Johnson et al. (2016) tested Peruvian L2 English learners' form recall knowledge using the full PLT and adopted multiple scoring systems that either did or did not penalize for orthographic inaccuracy (strict and lenient scoring). Their study showed that test scores based on all scoring systems were significantly correlated with L2 writing scores, but the strongest correlation emerged for the strict scoring system. These findings support the importance of productive vocabulary knowledge in L2 writing performance, and they suggest that knowledge of the correct form is an important aspect of this relationship.

In conclusion, Section 2.4. illustrated that vocabulary knowledge seems to be a key predictor of both receptive and productive L2 skills. While the field has dedicated much attention to the role of receptive vocabulary knowledge in L2 skills, less is known about the role of productive vocabulary knowledge in speaking and writing. Existing research has indicated that productive vocabulary knowledge predicts L2 writing performance, and that it may be important to take into account correct orthography in the scoring system, but the evidence so far remains limited.

2.5. Vocabulary use and skills

In addition to studies on vocabulary knowledge and the four skills, L2 research has also investigated the role of vocabulary use in L2 speaking and writing skills as a proxy for language proficiency. As explained in Section 2.3.1., vocabulary use is most commonly operationalized as lexical richness, comprised of subconstructs like lexical diversity and lexical

sophistication (e.g., Crossley, 2020; Kyle, 2020b; Read, 2000). In research on vocabulary use, speaking and writing skills are typically interpreted in terms of performance on a speaking or writing task, measured through holistic scores by human raters. The following subsections describe some existing research on the relationship between indices of lexical richness and speaking and writing.

2.5.1. Lexical diversity

Lexical diversity has repeatedly been successfully linked to L2 speaking and writing skills (e.g., Crossley & McNamara, 2012; Guo et al., 2013; Kyle et al., 2024; Noreillie et al., 2020). First, indices of lexical diversity have been shown to be directly associated with human ratings of L2 speaking performance (e.g., Bulté & Roothooft, 2020; Kyle et al., 2024; Kyle & Eguchi, 2023; Noreillie et al., 2020). Kyle and Eguchi (2023) investigated the relationship between MATTR and holistic speaking scores of Japanese L2 learners of English. Their study showed that lexical diversity was moderately correlated with holistic scores, and diversity was among the four indices explaining the most variance in the holistic speaking scores. Similar results were obtained by Noreillie et al. (2020), who investigated the relationship between HD-D and speaking scores of L2 learners of French at a low-intermediate level. They found that lexical diversity was the top predictor of holistic speaking scores.

Lexical diversity has also emerged as a predictor of L2 writing performance (e.g., Crossley & McNamara, 2012; Daller & Phelan, 2007; Guo et al., 2013; Kyle et al., 2021). Crossley and McNamara's (2012) study showed a moderate correlation between D and holistic writing scores (r = 0.43), and lexical diversity explained 18% of the variance in the holistic scores. In addition, in Kyle et al. (2021), human judgements of lexical diversity in L2 writing were found to be strongly associated with computational indices of lexical diversity (number of types, MATTR, HD-D, MTLD and MTLD-W), indicating that these indices provide a good representation of human raters' notion of lexical diversity.

Finally, research has also investigated the relationship between lexical diversity and L2 learners' proficiency level (e.g., Díez-Ortega & Kyle, 2024; Engber, 1995; Kisselev et al., 2025; Treffers-Daller et al., 2018; Zenker & Kyle, 2021). Treffers-Daller et al. (2018) showed that indices of lexical diversity (number of types, TTR, Guiraud's Index, D, HD-D and MTLD) were able to distinguish between L2 English learners of different proficiency levels according to the Common European Framework of Reference (CEFR; Council of Europe, 2020). In a study by Zenker and Kyle (2021), linear mixed-effects models showed that L2 English learners' CEFR level predicted their lexical diversity as measured through a wide range of

indices (TTR, Root TTR, Log TTR, Maas' Index, HD-D, MATTR, and three variations of MTLD). While lexical diversity in other L2s has been investigated to a lesser extent, the positive relationship between lexical diversity and L2 proficiency has also been demonstrated for L2 French (Treffers-Daller, 2013), L2 Spanish (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024), L2 Korean (Sung et al., 2024), and most recently L2 Russian (Kisselev et al., 2025). However, relatively few studies have employed MATTR as a measure of lexical diversity (though see e.g., Díez-Ortega & Kyle, 2024; Sung et al., 2024; Zenker & Kyle, 2021), so it remains to be confirmed to what extent MATTR can distinguish between L2 learners in L2 English and L2 French.

2.5.2. Lexical sophistication

Lexical sophistication, the use of sophisticated or advanced words, is also positively related to L2 speaking and writing skills, although the findings have not always been conclusive. Several studies have found lexical sophistication, measured through word frequency, to correlate with holistic ratings of spoken L2 output (Crossley & McNamara, 2013; Kyle & Crossley, 2015; Noreillie et al., 2020). For instance, a study by Crossley and McNamara (2013) demonstrated that word frequency (along with the number of types) explained 61% of the variance in holistic ratings of speaking performance in L2 English. However, some other studies have observed no significant correlations between the use of low-frequency words and holistic ratings of speaking performance in L2 English (e.g., Lu, 2012; Uchihara & Clenton, 2020). This discrepancy in research findings could be partly explained by the fact that spoken texts tend to consist of fewer low-frequency words than written texts (Nation, 2006).

With regard to written texts, frequency-based lexical sophistication has more consistently emerged as a predictor of holistic ratings of L2 output, although studies have mainly focused on L2 English (e.g., Crossley et al., 2011; Crossley & McNamara, 2012; Daller & Phelan, 2007; Guo et al., 2013; Johnson et al., 2016; Kyle et al., 2018; McNamara et al., 2010; Zhang, 2022). Daller and Phelan (2007) found that holistic ratings of written L2 texts were predicted by the use of low-frequency words, as measured through the proportion of advanced types and the proportion of words beyond the 1,000 most frequent word families (P_Lex; Meara & Miralpeix, 2016). In addition, the average word frequency (mean-based rather than band-based approach) of L2 English writing has also been shown to predict ratings of L2 essays. This has been found for average word frequency based on all words (Kyle et al., 2018; McNamara et al., 2010; Zhang, 2022) as well as only content words (Guo et al., 2013; Kyle et al., 2018), although Zhang

(2022) has suggested that the relationships may differ across writing genres.

Several studies have also found low-frequency vocabulary use to be a predictor of L2 proficiency, in L2 English (Crossley et al., 2013; Daller & Xue, 2007; Laufer & Nation, 1995), L2 Spanish (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024; Tracy-Ventura, 2017), and L2 French (Forsberg-Lundell & Lindqvist, 2012; Lindqvist, 2010; Ovtcharov et al., 2006). However, not all studies have established an association between L2 proficiency and the use of sophisticated words in English (Henriksen & Danelund, 2015; Laufer, 1998; Levitzky-Aviad & Laufer, 2013). These divergent findings may be partly explained by the study design and the sample, since the latter studies investigated cross-sectional differences between grades in secondary or tertiary education, rather than differences between learner groups categorized into different proficiency levels. Another potential reason for the lack of a relationship in the latter studies could be related to lexical sophistication being measured through the proportion of low-frequency words (see also Section 2.3.1. for a more detailed discussion of band-based approaches to lexical sophistication). The inconclusive findings on the use of low-frequency words in L2 English warrant further research into lexical sophistication across grades in secondary school. Moreover, it remains to be seen to what extent these findings may extend to L2 French.

Given the widely acknowledged importance of word frequency in L2 vocabulary development (e.g., Consolini & Kyle, 2024; Ellis, 2002; Kyle et al., 2018; Nation, 2006), the present thesis focuses on frequency-based lexical sophistication. As noted in Section 2.3.1., it should be mentioned that in addition to word frequency, lexical sophistication is increasingly interpreted as a multi-dimensional construct related to multiple word properties, such as psycholinguistic word information and multi-word units (e.g., Crossley, 2020; Kim et al., 2018; Kyle, 2020b; Nicklin et al., 2025). For instance, bigram strength of association, the likelihood of two words co-occurring (chronologically or syntactically), seems to be a strong predictor of L2 speaking performance (e.g., Kyle & Eguchi, 2023), L2 writing performance (e.g., Bestgen & Granger, 2014), and L2 proficiency level (e.g., Consolini & Kyle, 2024). In Crossley et al. (2015), collocation accuracy accounted for more than 80% of the variance in holistic ratings of L2 speech and writing. However, indices based on multi-word units target lexico-grammatical patterns rather than single word use, and thus, arguably reflect *phraseological* complexity rather than *lexical* complexity (see e.g., Bestgen & Granger, 2014; Paquot, 2019; Vandeweerd et al., 2022). They will therefore not be further discussed in this thesis.

In summary, Section 2.5. showed that lexical diversity and lexical sophistication have emerged as important predictors of L2 writing performance. More lexically diverse texts have consistently been shown to receive higher holistic ratings, and more proficient learners tend to use more varied vocabulary, although research using MATTR as an index is relatively limited. In terms of lexical sophistication, previous findings have been somewhat inconsistent. While the use of less frequent words seems to be related to higher holistic writing scores, it has not been consistently linked to L2 proficiency, and few studies to date have combined a band-based and mean-based approach for a more detailed account of lexical sophistication.

PART III: Vocabulary learning

2.6. Theoretical accounts of vocabulary learning

This section discusses theoretical underpinnings of vocabulary development related to L2 exposure and input. The first subsection will focus on usage-based approaches to L2 learning, followed by the second subsection which will elaborate on the input hypothesis.

2.6.1. Usage-based learning

The field of SLA has been marked by an increasingly solidified usage-based perspective on L2 learning (Ortega, 2013a). Usage-based accounts of language learning, often drawn from cognitive psychology research, are centered around the notion that languages are learned through repeated exposure and use (e.g., Ellis & Collins, 2009; Ellis & Wulff, 2020; Tomasello, 2003). When L2 learners encounter words in L2 utterances, they start to form lexical representations of those words in their mental lexicon. With repeated and frequent exposure to those words, the strength of those representations increases, making them more entrenched (Harris, 1998; Langacker, 1987). Stronger representations tend to be more easily accessible and readily available for learners (Conklin, 2020; Ellis & Wulff, 2020; Puimège, 2024), which facilitates understanding of L2 input (receptive) and use in learners' own L2 output (productive). As a result, language learners tend to have better knowledge of higher-frequency words than lower-frequency words, as they are more frequently exposed to frequent words (e.g., Milton, 2009; Schmitt, 2010a).

These so-called 'frequency effects' have been widely attested for (see e.g., Divjak & Caldwell-Harris, 2015; Ellis, 2002; Libben, 2022). Some words in a language occur more frequently than others (e.g., one is

generally more likely to encounter the word *brain* than the word *cerebrum*), and humans seem to have a good sense of these frequency differences (Ellis, 2002). Shapiro (1969) asked several groups of people (e.g., primary and secondary school pupils, college sophomores, adult professionals) to estimate the relative frequencies of English words. The study showed that the estimated frequency rankings were highly similar across the groups, and they were strongly correlated with the actual frequency rankings.

This intuitive sense of word frequency is based on people's experience with and use of the language. Through extensive contact with the language, humans acquire statistical information about the language, including the probability of occurrence of certain words and constructions (Ellis & Wulff, 2020; Onnis, 2012; Rebuschat & Williams, 2012). These probabilities are learned through humans' personal experiences with language, and this leads to the emergence of a mental language system containing information about the statistical regularities of language (Bybee, 1998; Ellis, 2002; Ellis & Wulff, 2020; Ortega, 2013b). It should be noted that probability of occurrence can be context-independent (simple probability), or it can be shaped by the specific context of the communicative situation (conditional probability) (Shukla et al., 2012). For instance, the word *cerebrum* may have lower simple probability than *brain*, but in a neuroanatomy textbook, *cerebrum* may have higher conditional probability. In fact, studies have indicated that lexical decision times can be dependent on participants' social group or professional area of expertise (e.g., Divjak & Caldwell-Harris, 2015), providing support for the notion that frequency effects are shaped by people's experience with language.

In addition to the effect of frequency on the strength of lexical representations, frequency also affects ease of access to these representations (e.g., Ellis & Wulff, 2020; Libben, 2022; Nahatame & Uchida, 2025; Puimège, 2024). Models like those by Forster (1976) and Morton (1969) provide theoretical accounts for these frequency effects on lexical processing during word recognition. Both of these models postulate that lexical representations of higher-frequency words are activated before less frequent words[10]. Frequency is, thus, related to how quickly and efficiently a word can be activated or accessed. This is supported by studies showing that high-frequency English words are processed more quickly during reading (e.g., Nahatame & Uchida, 2025), during lexical decision tasks (e.g., Chen et al., 2018; Whaley, 1978) and during meaning recognition tests (e.g., Laufer & Nation, 2001). These frequency effects have recently been shown to hold across twelve alphabetic languages (Kuperman et al.,

[10] Morton's (1969) logogen model accounts for this through the notion of an activation threshold based on a word's probability of occurrence, so that more frequent words are activated before less frequent words. Forster's (1976) model postulates that more frequent words are accessed before less frequent words through a serial search of the mental lexicon.

2024), and they seem to be stronger in the L2 than in the L1 (Durrant et al., 2022).

While the accounts for frequency effects are generally based on lexical processing during word recognition, similar processes may be at work during word production. Studies have found that the lexical frequency effect extends to speech and written production, whereby less frequent words are produced more slowly than more frequent words in picture naming tasks (e.g., Almeida et al., 2007; Bonin et al., 1998; Cuetos et al., 2006; Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1965). Based on the presumptions that (a) humans have a mental lexicon containing statistical information about the frequency of words, (b) word frequency impacts the strength of lexical representations, and (c) stronger representations are more easily accessible, it seems logical that frequency effects also apply to production. For instance, the word *brain* is generally more frequent than *cerebrum* (a), which means it likely has a stronger representation in the mental lexicon (b), causing it to be accessed more efficiently during retrieval (c).

In terms of vocabulary use in speaking and writing, this seems to align well with functionalist meaning-to-form approaches to L2 production. These approaches depart from the basic tenet that written and spoken production require L2 learners to express concepts (meanings) they already know in their L1, and they need lexical devices (forms) in the L2 in order to do so (Bardovi-Harlig, 2020). Lexical access, then, is the process of mapping those concepts onto entries in the mental lexicon (Levelt et al., 1999; Roelofs, 1997). This can sometimes be a one-to-one mapping, but often the multifunctionality principle is at play: multiple forms can be used to express a single meaning (Bardovi-Harlig, 2020). The example of *brain* and *cerebrum* illustrates this well. Based on frequency effects, it can be expected that in order to express the concept of 'an organ that controls bodily functions', the average L2 learner is more likely to use the word *brain* than the word *cerebrum*. The likelihood of *cerebrum* being used can be expected to increase along with a learner's experience with the L2, although this likelihood is also dependent on the contexts to which the learner is exposed (compare e.g., a student in the field of linguistics and a student in the field of medicine).

2.6.2. Input hypothesis

Usage-based approaches emphasize the role of learners' exposure to the L2 for language learning, and the importance of exposure and input for language learning has long been acknowledged. Stephen Krashen introduced his theory of second language acquisition to the field of SLA in the 1980s (Krashen, 1982). A key component of Krashen's theory is the

input hypothesis, which states that L2s are acquired through meaningful, comprehensible input. More concretely, according to the input hypothesis, L2 learners need to be exposed to L2 input while being focused on the meaning (not the form). The input should be slightly above their current level in terms of linguistic forms, yet still understandable by means of context and knowledge of the world. Krashen argues that through comprehensible input, L2 productive ability 'emerges' over time (p. 22) and it cannot be taught directly.

The input hypothesis is based on another central assumption of Krashen's theory, namely, the acquisition–learning distinction. Acquisition is defined as implicit and subconscious L2 development, and learning is defined as explicit and conscious knowledge of L2 forms[11]. According to Krashen, L2s are acquired mainly subconsciously, similar to how children acquire their L1. Since knowledge is assumed to be acquired through comprehensible input rather than a focus on linguistic forms, most L2 utterances are produced based on acquired rather than learned knowledge. Explicit, learned knowledge can serve merely as a 'monitor' to edit utterances before or after they are produced in speaking and writing.

The core idea behind Krashen's input hypothesis, the importance of L2 exposure for L2 learning, is widely accepted and even a fundamental aspect of usage-based approaches to L2 learning. However, Krashen's notion of 'comprehensible input' has been criticized for being too restrictive. First, the input hypothesis postulates that L2s can only be acquired from input. This is in contrast with usage-based approaches which have long defined 'usage' as more than only input (Wulff, 2021), accounting for learners' contact with the L2 through input (understanding the L2 through listening and reading) as well as output (producing the L2 in speech and writing). This is in keeping with Swain's (1985) output hypothesis, which states that learning can also occur through attempts to produce comprehensible output that is 'conveyed precisely, coherently, and appropriately' (Swain, 2005, p. 473). Second, the input hypothesis claims that input should be focused on meaning and not on form. This is not in line with research that has shown the value of attention to linguistic forms (e.g., Pellicer-Sánchez & Montero Perez, 2024; Schmidt, 1990) and the benefits of word-focused instruction (e.g., Laufer, 2020; Loewen, 2020; Nation & Webb, 2011).

[11] Krashen's acquisition–learning distinction will not be adopted in the present thesis, because the data described in this thesis do not allow us to clearly identify which L2 knowledge the participants have either acquired or learned. The terms *acquisition* and *learning* will thus be used interchangeably (except when describing Krashen's ideas), regardless of whether they refer to conscious or unconscious cognitive processes (see also Ellis, 2004; Ortega, 2013b).

In short, Section 2.6. outlined theoretical accounts of vocabulary learning. Usage-based accounts of vocabulary learning provide a useful framework for the frequency effects that have been widely established in (L1 and) L2 research. These frequency effects can be explained in terms of the degree of exposure to and use of L2 vocabulary, whereby more frequent vocabulary is more strongly represented in the mental lexicon and, thus, more easily accessible during recognition and production. The usage-based framework resonates well with Krashen's input hypothesis, although the question arises whether the notion of comprehensible input is too restrictive to represent vocabulary learning from exposure and use.

2.7. Instructed learning

A traditional and common way to learn an L2 is through formal, in-class instruction. The subsections below will elaborate on empirical research investigating how much vocabulary can be learned throughout formal instruction, as well as the ways in which vocabulary can be addressed in an instructional setting.

2.7.1. Instruction and L2 vocabulary learning

There is a body of research that has investigated L2 learners' vocabulary knowledge throughout different stages of instruction, mainly using one-off studies or cross-sectional designs. Although there can be large individual variation in learners' vocabulary knowledge (e.g., Miralpeix, 2020; Miralpeix & Muñoz, 2018; Orosz, 2009; Takala, 1985), vocabulary knowledge is generally known to increase with learners' proficiency level (Miralpeix, 2020; Waring, 1997; Zareva et al., 2005) and with learners' length of formal instruction (Milton, 2009; Orosz, 2009; Peters, Noreillie, et al., 2019). Most research on the development of vocabulary knowledge has focused on recognition knowledge (e.g., Schmitt, 2020).

First, some studies have measured vocabulary knowledge at the level of form recognition (e.g., Miralpeix, 2020; Miralpeix & Muñoz, 2018; Orosz, 2009), and they indicate that form recognition knowledge increases gradually throughout formal instruction. Spanish university students of L2 English at an intermediate level have been found to recognize word forms for approximately 4,000 word families, while those at an advanced level could recognize approximately 6,000 word families (Miralpeix, 2007, as cited in Miralpeix, 2020). In Miralpeix and Muñoz's (2018) study, upper-intermediate L2 English learners could recognize on average 5,000 word families, although the participants' form recognition knowledge ranged from 2,500 to 7,200 word families. A study with Hungarian L2 English learners showed that third-grade (primary school) students had a form

recognition vocabulary size of 350 word families, while sixth-grade students had developed a vocabulary size of almost 1500 word families (Orosz, 2009). These studies show that L2 vocabulary knowledge at the level of form recognition increases throughout formal instruction.

Other studies have measured vocabulary knowledge at the level of meaning recognition, usually with samples of university students. In a study with Chinese L2 English learners in the first year of university (X. Zhang & Lu, 2014), participants demonstrated mastery of the 2,000 most frequent word families (i.e., a score of at least 26/30 for that frequency level). Similarly, Fan (2000) established an average score of 86% for the 2,000 frequency level for Chinese university students, although this average score dropped considerably to 67% for the 3,000 level. These numbers seem to be in line with a recent study by Lin (2023), which showed an average score of 70% on the 2,000 and 3,000 level taken together. University students in Germany have been found to perform slightly better, with average scores of 93% for the 2,000 level and 78% for 3,000 (Tschirner, 2004). In contrast, Webb and Chang (2012) found that L2 English learners in Taiwan had relatively limited meaning recognition knowledge. By their final year of university, 73% of students in a language-oriented program demonstrated mastery of the 1,000 most frequent word families and only 33% had mastered the 2,000 level. For students in a non-language-oriented program, only 33% had mastered the 1,000 level, and this number dropped to 6% for the 2,000 level.

Some L2 vocabulary research has adopted a longitudinal design to model vocabulary development in the same learners over time. De Wilde et al. (2021) tested primary school children on their L2 English meaning recognition knowledge in grade 6 (last year of primary school) and again in grade 8 (second year of secondary school). The authors found that learners' meaning recognition knowledge increased, with the average test score rising from 69% to 81%. With university students, Zhang and Lu (2014) indicated that meaning recognition knowledge increased for all frequency levels on the VLT at each time point. Participants already demonstrated mastery of the 2,000 level at time 1, they also mastered the 3,000 level by time 2, and they increased their knowledge at the other frequency levels. Another longitudinal study of meaning recognition knowledge showed that university-level learners of L2 English gained knowledge of an average of 200 to 300 word families per year (Webb & Chang, 2012). At the level of form recognition, beginner L2 English learners have been shown to gain knowledge of 500 word families on average per year, albeit with large individual variation (Milton, 2009), whereas average gains of 300 to 400 words per year have also been found for L2 English learners from grade 3 to grade 12 (Orosz, 2009).

Research on vocabulary development at the recall level is more limited, although recall knowledge also seems to improve with increased formal instruction. Some studies have adopted translation test formats in order to test L2 English vocabulary knowledge at the level of both meaning recall (L2-to-L1 translations) and form recall (L1-to-L2 translations) (e.g., Takala, 1985; Webb, 2008). In keeping with the incremental nature of vocabulary knowledge (e.g., Laufer & Goldstein, 2004), L2 learners typically have better meaning recall than form recall knowledge. In Webb (2008), Japanese learners of L2 English demonstrated mastery (a score of 26/30 or higher) of items up to the 3,400 most frequent words for meaning recall, while they could only demonstrate mastery up until the 1,900 most frequent words for form recall. With a strict scoring system where spelling errors were penalized, the learners were not found to master the highest frequency level (701 to 1,900 most frequent words)[12]. Other studies using form recall tests in the PLT format, have also found that vocabulary knowledge at the form recall level was less developed than at the meaning recognition level (e.g., Fan, 2000; Tschirner, 2004; Waring, 1997). In terms of known word families, Laufer (1998) found that Hebrew secondary school learners of L2 English had form recall knowledge of 1,700 word families in grade 10 (a test score of 30%) and 2,550 word families in grade 11 (test score of 46%), compared to 1,900 and 3,500 word families at the level of meaning recognition. While the cross-sectional differences in form recall knowledge appeared to be smaller than in form and meaning recognition knowledge, there was still improvement with increased length of formal instruction.

Only few studies have investigated vocabulary development in other languages than L2 English, with some exceptions (Batista & Horst, 2016; Eyckmans, 2004; Milton, 2006; Noreillie et al., 2018). One exception is Batista and Horst (2016), who investigated meaning recognition knowledge of L2 French learners at four proficiency levels. In line with research on L2 English, they found that test scores decreased with each frequency level, and higher-proficiency groups outperformed the lower-proficiency groups. While the most advanced learner group was found to have a vocabulary size of about 6900 word families (average test score of 77%), the beginner group had a vocabulary size of 2700 word families (average test score of 32%). In a comparison of meaning recognition and form recall knowledge of L2 French secondary school students, Noreillie et al. (2020) found that test scores were higher on the meaning recognition test (mastery up to the 2K level; on average 92% for the 1K level, 87% for 2K) than on the form recall test (no mastery of 1K or 2K level; on average 72% for 1K, 62% for 2K). This seems to support the implicational nature of vocabulary knowledge that has been observed for L2 English (e.g.,

[12] Webb (2008) uses 'words' to refer to the frequency levels of the target items instead of a more specific counting unit.

González-Fernández, 2025; González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004).

Even fewer studies have compared vocabulary development of the same individuals across multiple L2s. An exception is Peters, Noreillie, et al. (2019), who adopted a cross-linguistic and cross-sectional design to compare meaning recognition knowledge in L2 English and L2 French of Flemish learners in grade 8, grade 10, and the first year of university. Their study showed that the group with the most years of formal instruction (university) outperformed the second group (grade 10), who in turn outperformed the group with the lowest length of instruction (grade 8). While this pattern was the same for both English and French, there was a large discrepancy in the scores between the L2s, with the English test scores being much higher (47% in grade 8, 67% in grade 10, 82% at university) than the French test scores (24% in grade 8, 45% in grade 10, 63% at university). This seems counterintuitive, given that the learners' length of instruction was much larger for French than English, as French instruction starts before English instruction in Flanders. This discrepancy between English and French has been explained in light of the large amounts of out-of-school contact with L2 English and the more positive attitudes towards English of L2 learners in Flanders (e.g., Dewaele, 2005b; Housen et al., 2001; Peters, Noreillie, et al., 2019). The findings of this cross-linguistic study provide evidence that results for one L2 cannot simply be generalized to other L2s, especially when the learning context may differ between the L2s. Given the paucity of cross-linguistic studies into vocabulary development, especially at a productive level of mastery, more research is needed.

2.7.2. Word-focused instruction

For many sequential, non-immersion learners[13], formal instruction is the primary and sometimes the sole source of L2 input (though see Section 2.8.). Since classroom time is usually limited to a few hours per week, formal instruction has been characterized by two key elements: (a) the selection of specific features to be addressed, and (b) specific attention to formal characteristics (Ellis, 2004). In the context of vocabulary instruction, this often implies word-focused instruction, which can involve practicing with decontextualized vocabulary as well as working with vocabulary during communicative opportunities (Laufer, 2014). According to Nation's (2007) four strands, a formal language course should promote L2 learning by means of the four following types of activities: (1) meaning-

[13] 'Sequential, non-immersion learners' refers to learners in a context in which (a) the L2 is learned after the L1 has already been acquired, and (b) the L2 is not learned in a naturalistic setting, as is the case in migration contexts, for instance.

focused input, through listening and reading; (2) meaning-focused output, through speaking and writing; (3) language-focused learning, through explicit teaching of vocabulary; and (4) fluency development, aimed at increasing automaticity. The four strands provide a framework that integrates meaning- and form-focused activities as well as receptive and productive aspects of L2 learning.

For meaning-focused input, Nation (2007) emphasizes the importance of (a) large quantities of input, (b) a large portion of already familiar vocabulary in the input, and (c) learners' motivation to understand the input. This is in line with what is known about the role of input (e.g., Krashen, 1982), lexical coverage (e.g., Hu & Nation, 2000; Schmitt et al., 2011), and motivation (e.g., Dörnyei, 2005) in L2 learning. A recent meta-analysis showed that, indeed, vocabulary can be learned from meaning-focused activities (Webb et al., 2023). The second strand, meaning-focused output, highlights the importance of producing the L2 in speaking or writing, in keeping with Swain's (1985) output hypothesis. Meaning-focused output refers to writing or speaking tasks that are mainly focused on conveying a message, that is, they prioritize the expression of meaning over a focus on language features (see also Ellis, 2017; Van den Branden, 2022, for discussions of task-based language education). This primary focus on meaning is in contrast with the third strand, language-focused learning, which involves deliberate attention to language features, as is commonly the case in form- or word-focused instruction[14] (see Webb et al., 2020, for a meta-analysis of intentional vocabulary learning activities). For language-focused learning, Nation underscores the importance of deliberate, repeated attention to language features, in line with incidental vocabulary research that has demonstrated positive effects of repetition and attention on vocabulary learning (Conklin et al., 2018; De Vos et al., 2018; Webb et al., 2023). The final strand is fluency development, which is aimed at helping learners to 'make the best use of what they know' (Nation, 2007, p. 7). It is characterized by activities with known words only, with some pressure to perform faster than usual (e.g., speed reading, 10-minute writing).

An important aspect of the language-focused strand (Nation, 2007) is the fact that learners' deliberate attention is repeatedly drawn to linguistic features. This is in keeping with Schmidt's (1990) noticing hypothesis, which posits that learning is most likely to occur when there is attention to linguistic forms, that is, when learners consciously detect the presence of a linguistic form (regardless of their awareness of any underlying rules or patterns) (Leow, 2019; Loewen, 2020; Schmidt, 1990).

[14] Nation (2007) notes that terms like 'focus-on-form' and 'form-focused' instruction may be misleading because language-focused learning can also involve a deliberate focus on meaning.

Noticing can be affected by a range of word- and learner-related variables like salience (how much a form stands out), frequency (how often a form occurs) and individual differences (e.g., motivation). Some potential ways to enhance noticing in an instructional context are to provide glosses (e.g., Abraham, 2008; Hulstijn & Laufer, 2001), to let learners use a dictionary to search the meaning of a word in a reading task (e.g., Laufer, 2020), and to use textual enhancement, for instance, through bolding or underlining (e.g., Vu & Peters, 2022). It is also important to ensure repeated exposure, for instance, by recycling vocabulary across tasks to allow for spaced repetition (e.g., Nation, 2007), although some studies have shown that vocabulary is often not effectively recycled in L2 textbooks (Bergström et al., 2025; Nguyen, 2021; Van Parys et al., 2024).

Furthermore, scholars seem to agree that learners can benefit from actively engaging with vocabulary. Schmitt (2008, p. 339) refers to this with his concept of *engagement*, relating to 'virtually anything that leads to more exposure, attention, manipulation, or time spent on lexical items', which all seem to contribute to more optimal learning conditions. Existing frameworks to evaluate task effectiveness, like the Involvement Load Hypothesis (Laufer & Hulstijn, 2001), the Involvement Load Hypothesis Plus (Yanagisawa & Webb, 2022), and the Technique Feature Analysis (Nation & Webb, 2011), seem to have a shared acknowledgement of the role of active engagement in vocabulary learning. This aligns well with the depth of processing theory (Craik & Lockhart, 1972), which states that words which are processed more deeply, are more likely to be retained. Indeed, studies have found that word production exercises after a reading task benefitted word retention more than reading-only (e.g., Eckerth & Tavakoli, 2012; Jin & Webb, 2025; Laufer & Rozovski-Roitblat, 2015). Laufer & Rozovski-Roitblat (2015) demonstrated that learners who encountered target words once in a reading text and then once or twice in vocabulary exercises (2–3 exposures) had higher learning gains than learners who encountered the target words 18–21 times in a reading text.

There is some empirical evidence that word-focused activities in the L2 classroom may result in better word retention than word-unfocused activities (Laufer, 2003, 2020; Webb et al., 2023). A word-focused approach allows teachers to directly provide learners with lexical information which they may not easily pick up from input beyond the instructional setting, especially if input outside of the L2 classroom is limited. Laufer (2020, p. 362) gives the example of expanding learners' knowledge with a new structure in which a phrasal verb can occur (e.g., 'providing something for someone' and 'providing someone with something'), or with certain nuances in word meaning and use (e.g., 'fee' and 'payment'). Such additional lexical information in instruction may help learners to expand their knowledge of vocabulary. This positive impact of instruction seems to

be supported by Laufer and Paribakht's (1998) findings on the receptive and productive vocabulary knowledge of learners of English as a foreign language (EFL; learners in Israel) and learners of English as a second language (ESL; learners in Canada). The study showed that EFL learners demonstrated a smaller discrepancy between receptive and productive vocabulary knowledge than the ESL learners, indicating that although ESL learners had a larger vocabulary size, the EFL learners had better productive knowledge of the words they knew receptively. This may suggest that the extensive input in naturalistic settings is beneficial to the development of receptive vocabulary breadth, and the type of input in instructional settings may lend itself particularly well to fostering higher levels of mastery.

Examples of commonly used word-focused activities in the L2 classroom include word cards (i.e., creating flashcards with the L2 word and an L1 translation), sentence writing (i.e., producing a sentence with the L2 word), fill-in-the-gap for isolated sentences (i.e., completing missing L2 words in unconnected sentences), reading with fill-in-the-gap for a text (i.e., completing missing L2 words in a text based on a list of [too many] options with L2 words, L2 explanations, L1 translations and sample sentences), finding words in a text (i.e., after reading, searching the text for several provided meanings), and rewording sentences (i.e., using a different word to rephrase given sentences) (e.g., Laufer, 2020; Loewen, 2020; Nation & Webb, 2011). A meta-analysis by Webb et al. (2020) compared learning gains at the level of meaning and form recall from four common activities: flashcards, word lists, writing, and fill-in-the-blanks. They found that flashcards and word lists seemed to result in the largest immediate learning gains (60% on meaning recall tests and 58% on form recall tests), followed by writing tasks and fill-in-the-blanks. However, long-term retention was shown to be much lower through delayed posttests, with learning gains decreasing from 60% to 39% for meaning recall tests, and from 58% to 25% for form recall tests.

Since classroom time is limited, 'a manageable quantity of the most important vocabulary for learners' needs' has to be selected (Laufer, 2014, p. 38). An empirical study by He and Godfroid (2019) has attempted to specify which type of words should ideally be prioritized in order to maximize the limited L2 classroom time. Given the cost-benefit principle, which entails that learners should get the best return for the effort they put in, He and Godfroid (2019) recommend focusing on words that have high frequency, high usefulness and low difficulty. These criteria were proposed by Laufer and Nation (2012) because they are related to how easy the words are to learn and how likely learners are to frequently encounter and use them. Words with mid-level frequency, mid-level usefulness and high difficulty should also be addressed for the purpose of solidifying learners'

progress, and should be given priority over words with low frequency, low usefulness and mid-level difficulty. Finally, He and Godfroid (2019) recommend replacing words with low frequency, low usefulness and high difficulty with more frequent and useful academic vocabulary.

In brief, Section 2.7. indicated that formal instruction can contribute to L2 vocabulary development. Instruction can support, expand and solidify learners' vocabulary knowledge through activities that encompass repeated exposure and use of selected vocabulary, ideally with high levels of engagement and attention, and preferably for vocabulary that provides a good return on investment for the learners. However, existing research on the impact of formal instruction has often focused on recognition knowledge of L2 English, most commonly with university students. Little is known about the development of recall knowledge throughout secondary education and particularly in other L2s than English, like French.

2.8. Out-of-school learning

Apart from formal instructional contexts, L2 learning can also take place outside of the L2 classroom. The first subsection below discusses the terminology that has been used to refer to this phenomenon and attempts to define the notion of out-of-school contact. The following subsections address existing research on the role of out-of-school contact and specific activities in L2 learning.

2.8.1. Defining out-of-school contact

Since the amount of L2 input in a formal instructional setting is typically limited, L2 learners can seek additional input beyond the L2 classroom (e.g., Amengual-Pizarro & Alonso Alonso, 2024; Laufer, 2019; Loewen, 2020). A recent review by Sundqvist (2024) highlighted that for many L2 learners of English in (Northern) European contexts, out-of-school language contact has replaced formal instruction as the starting point for learning English. In fact, prior to formal instruction, young learners of L2 English can already have receptive and productive vocabulary knowledge of some high-frequency words (Bollansée et al., 2021; De Wilde et al., 2022; Puimège & Peters, 2019a).

Out-of-school language contact has gained much traction in recent decades, and many terms have been proposed to refer to the notion of L2 contact outside of the school walls. One widely used term is 'extramural exposure' (Sundqvist, 2009; Sylvén, 2006), also commonly referred to as 'extramural English' (EE) due to the dominant focus on English out-of-school contact in existing research (Kusyk et al., 2025; Sundqvist, 2024). Various other terms have been coined to describe L2 learning beyond the

L2 classroom, each with their own focus. Some terms have emphasized the often incidental and informal nature of out-of-school learning, like 'incidental foreign language acquisition' (Kuppens, 2010), 'unintentional learning of English' (Forsman, 2004), and 'informal second language learning' (Arndt, 2019). Other terms have highlighted the fact that the L2 contact occurs beyond the formal instructional context, such as 'language learning beyond the classroom' (Benson & Reinders, 2011), 'out-of-class learning' (Benson, 2013), and 'language learning in the wild' (Wagner, 2015). Some terms also include specifications about the platform or medium where the L2 contact occurs, like 'online informal learning of English' (Sockett, 2014), 'out-of-class language learning with technology' (Lai & Gu, 2011), 'digital learning' (Niu et al., 2022), and 'informal digital learning of English' (Lee & Dressman, 2018). The four latter terms are closely related to the fields of computer-assisted and mobile-assisted language learning (CALL and MALL), for instance, when learners engage with the L2 through social media use or apps like Duolingo.

In spite of the wide range of terminology, researchers are typically referring to the same general phenomenon. The main idea behind 'out-of-school', 'extramural', or 'informal' exposure is that the learner comes into contact with the L2 outside of the L2 classroom, for instance, through watching TV, playing videogames, reading, watching YouTube videos, listening to music, or using social media. Following Sundqvist's (2009) interpretation, this L2 contact need not be initiated intentionally, nor for the deliberate purpose of learning, although it is possible that these intentions are present. For instance, one L2 learner may watch a TV show simply for the purpose of watching the show (no intention to learn), whereas another L2 learner may watch a TV show for the deliberate purpose of increasing their vocabulary knowledge (intention to learn). An unplanned L2 conversation with a stranger can also be classified as (unintentional) out-of-school contact. Sundqvist's definition, thus, takes a neutral position towards the L2 learner's intentions (see also Arndt, 2019). A final aspect of Sundqvist's (2009) interpretation is the autonomy of the learner in initiating the L2 contact. This implies that activities imposed by the L2 teacher (e.g., reading a book or watching a film as homework for class) do not fall under the definition of extramural exposure, even if these activities take place outside of the L2 classroom. Conversely, and perhaps paradoxically, if a learner watches an L2 YouTube video in class while they are expected to work on a different task, this can also be interpreted as out-of-school L2 contact in spite of the in-class location (Sundqvist, 2024).

The present thesis closely follows Sundqvist's (2009) interpretation of out-of-school L2 contact. It is defined as any learner-initiated L2 activity outside of or independent from the L2 classroom (i.e., it is not part of formal in-class instruction, and not initiated or imposed by

the teacher), regardless of any potential deliberate intention for learning, and regardless of the platform, medium or modality (online, offline; written, spoken or audiovisual). To refer to this phenomenon in the rest of this dissertation, the term 'out-of-school contact' will be used for three main reasons. First, it does not imply any activity or passivity on the part of the learner during their engagement with the L2 (unlike, for instance, 'extramural exposure', which seems to imply a certain passivity or seems to refer to input-based contact only). Second, the term does not suggest that L2 contact necessarily results in learning, unlike the range of terms comprising the word 'acquisition' or 'learning' (e.g., 'incidental foreign language acquisition', 'informal second language learning', 'out-of-class learning'). Third and finally, the term 'out-of-school contact' does not inherently refer to any specific L2 (unlike, for instance, 'extramural English' or 'informal digital learning of English').

2.8.2. Out-of-school contact and L2 vocabulary learning

The potential benefits of out-of-school contact for L2 learning correspond well with the theoretical framework of usage-based learning and the input hypothesis (see Section 2.6.). Given the importance of L2 input for L2 learning (e.g., Ellis & Collins, 2009; Krashen, 1982), out-of-school contact seems to be a useful way for L2 learners to increase their exposure to the L2 (although in contrast with the input hypothesis, this exposure is not necessarily input-based only). In addition, since out-of-school contact is learner-initiated, learners can choose independently how often, how long and in which ways to engage with the L2 in their spare time. Many of the activities are thus more 'leisure-oriented' than 'learning-oriented' (Lyrigkou, 2023; Schwarz & Ghamarian-Krenn, 2024), which implies that learners likely enjoy the activities they choose to engage in (e.g., Alonso Alonso, 2022; Sundqvist & Sylvén, 2014). More motivated learners may also be more likely to seek out the L2, which could result in both more extensive and more positively appraised L2 contact. Motivation and enjoyment have repeatedly been established as affective variables beneficial to L2 learning (e.g., Al-Hoorie, 2018; Dewaele et al., 2022; Dörnyei, 2005; Leona et al., 2021).

Research has indeed quite consistently demonstrated the potential benefits of out-of-school contact for L2 learning. Most research has focused on the impact of out-of-school contact in relation to receptive vocabulary knowledge, specifically in L2 English. Out-of-school L2 contact seems to be positively associated with English meaning recognition knowledge of primary school learners (e.g., De Wilde et al., 2020b; Gyllstad et al., 2025; Hannibal Jensen, 2017; Kuppens, 2010; Sylvén & Sundqvist, 2012), secondary school learners (e.g., Peters, Noreillie, et al., 2019; Sundqvist,

2019; Sundqvist & Wikström, 2015; Warnby, 2022), and university students (e.g., Busby, 2021; Ghamarian-Krenn, 2023; Peters, Noreillie, et al., 2019). A positive impact of out-of-school contact has also been found on form recognition knowledge (e.g., Schwarz, 2020) and meaning recall knowledge (e.g., Laufer & Vaisman, 2023). Some studies have suggested that out-of-school contact was a better predictor of receptive vocabulary knowledge than formal instruction (e.g., Busby, 2021; Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024; Warnby, 2022). In addition, out-of-school contact has also been linked to the receptive skills of reading (e.g., De Wilde et al., 2020b; Lindgren & Muñoz, 2013; Neagu et al., 2025; Sockett & Kusyk, 2015; Wouters et al., 2024) and listening (e.g., De Wilde et al., 2020b; Lindgren & Muñoz, 2013; Wouters et al., 2024).

Productive vocabulary knowledge and productive skills like speaking and writing have only received limited attention so far (Sundqvist, 2024). Some research has established positive correlations between out-of-school contact and form recall knowledge of single words (Bollansée et al., 2021; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012) as well as collocations (González-Fernández & Schmitt, 2015). However, there are also studies that have not consistently established a positive relationship between out-of-school L2 activities and form recall vocabulary knowledge (e.g., Ghamarian-Krenn, 2023; Kuppens, 2010). Further, findings on the relationship between out-of-school contact and speaking have provided a mixed picture. For instance, De Wilde et al. (2021) established clear correlations between out-of-school contact and speaking performance, while Lyrigkou (2023) showed that overall out-of-school contact did not predict the use of discourse markers in speaking. Lyrigkou (2023) did, however, establish a significant relationship with specific activities, like speaking/interacting and watching TV without subtitles. These findings seem to suggest that the relationship between out-of-school contact and productive skills is complex, and out-of-school contact may impact on receptive and productive vocabulary knowledge in different ways. Schwarz and Ghamarian-Krenn (2024) attribute this to the receptive nature of many popular out-of-school activities (e.g., listening to music, watching online videos, reading social media messages), which may be less effective at promoting productive vocabulary knowledge (see also e.g., Griffin, 1992; Webb, 2008, 2009).

In terms of vocabulary use, some initial efforts have been made to examine the relationship between out-of-school contact and vocabulary use in writing (e.g., Cobb & Horst, 2011; Kaatari et al., 2023; Olsson & Sylvén, 2015; Sundqvist, 2019). Two studies in Sweden have found a positive impact of out-of-school contact on lexical diversity in L2 English writing (Kaatari et al., 2023; Sundqvist, 2019). In contrast, out-of-school contact has not always been demonstrated to correlate with lexical

sophistication in L2 English writing (Cobb & Horst, 2011; Olsson & Sylvén, 2015). It is possible that out-of-school contact does not strongly impact on lexical sophistication because out-of-school activities tend to contain general and informal vocabulary, which may not encourage the use of more sophisticated words (Rød & Calafato, 2025). However, given the limited research on the relationship between out-of-school contact and written vocabulary use, more research is needed to clarify this relationship.

2.8.3. Out-of-school activities and L2 vocabulary learning

Out-of-school activities are typically incidental, since learners' main purpose is often not deliberate learning (although this purpose is not necessarily excluded; e.g., Lyrigkou, 2023; Schwarz & Ghamarian-Krenn, 2024; Sundqvist, 2024). In that sense, out-of-school research can be linked to incidental vocabulary research, which has provided ample evidence that vocabulary can be learned incidentally from activities like reading (e.g., Godfroid et al., 2018; Horst et al., 1998; Nation & Waring, 2024; Pellicer-Sánchez, 2016; Puimège et al., 2024) and watching audiovisual input (e.g., Kurokawa et al., 2024; Montero Perez et al., 2024; Muñoz et al., 2023; Pujadas & Muñoz, 2019; Rodgers, 2013; Wang & Pellicer-Sánchez, 2022). A meta-analysis of experimental studies (Webb et al., 2023) has shown that L2 participants learned on average 15–17% of target items through reading, and a smaller average of 5–7% through TV viewing. Krashen (2004) has referred to the learning gains through reading as 'the power of reading', stating that free voluntary reading is a powerful tool to promote L2 development.

However, this finding has not always been corroborated in research on out-of-school contact, where reading has not consistently been found to predict vocabulary knowledge (e.g., Arndt, 2023; Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019). An explanation for this discrepancy may lie in the fact that L2 learners do not seem to frequently engage in reading on their own initiative (e.g., De Wilde et al., 2020b; Laufer & Vaisman, 2023; Muñoz, 2020; Peters, Noreillie, et al., 2019; Zhang et al., 2021). This low L2 exposure through reading may not suffice for vocabulary learning, and limited variation in the data may also complicate the detection of a relationship between reading and vocabulary knowledge. Still, in some studies, out-of-school reading has been shown to correlate with meaning recognition knowledge (De Wilde, 2025; Peters, 2018), recognition and recall knowledge of the form-meaning link, homonymous/polysemous words, derivations and collocations (González-Fernández, 2021), classroom grades (Muñoz, 2020), reading comprehension (Neagu et al., 2025) and listening comprehension (Wouters et al., 2024).

A similar issue may be present for research findings related to listening to music. While music is a highly popular activity among L2 learners, it has not always been established as a positive predictor of receptive L2 vocabulary knowledge (Hannibal Jensen, 2017; Niitemaa, 2020; Peters, Noreillie, et al., 2019), listening comprehension (Muñoz & Cadierno, 2021) or receptive grammar knowledge (Muñoz et al., 2018; Muñoz & Cadierno, 2021). One explanation may be the limited variation in learners' reported engagement with music, because the majority of L2 learners listen to music frequently (De Wilde et al., 2020b; Lindgren & Muñoz, 2013; Peters, Noreillie, et al., 2019; Wouters et al., 2024). This interpretation is supported by Muñoz and Cadierno (2021), who found similar exposure to English-language music for two learner groups in more and less input-rich learning contexts (Denmark and Spain). Alternatively, the lack of a well-established relationship between music and L2 learning may reflect differences in how actively learners engage with music (Kerekes, 2015; Niitemaa, 2020; Sockett & Toffoli, 2012). For example, an intervention study by Pavia et al. (2019) showed that listening to songs in class with attention to the lyrics could benefit vocabulary learning. This may explain why some correlational studies have also detected a positive relationship between out-of-school music and receptive vocabulary knowledge (Puimège & Peters, 2019a) or listening and reading comprehension (Lindgren & Muñoz, 2013).

Watching TV, in contrast, has emerged as a common out-of-school L2 activity that is more consistently positively associated with receptive vocabulary knowledge (e.g., De Wilde et al., 2020b; Kuppens, 2010; Peters, Noreillie, et al., 2019) and, to a smaller degree, productive vocabulary knowledge (e.g., Bollansée et al., 2021). This is in line with experimental studies into incidental vocabulary learning from audiovisual input, which have shown the benefits of TV viewing for vocabulary learning (e.g., Avello & Muñoz, 2025; Peters & Webb, 2018; Rodgers, 2013; Webb, 2014). There are still questions, however, with regard to the benefits of non-subtitled, (L1-)subtitled or (L2-)captioned TV. Peters et al. (2016) found a consistent advantage for captioned TV compared to subtitled TV for adolescent L2 learners (15–18 years old), while Pujadas and Muñoz (2019) observed no differences in learning gains from captioned or subtitled TV with younger teenagers (13 years old), possibly due to a distinct impact of subtitles and captions depending on the proficiency level. In out-of-school research, results have also been mixed: (L1-)subtitled TV was found to have a positive impact on form recall translation scores of primary school learners in Kuppens (2010), whereas it had a negative correlation with form recall scores of primary school learners in Bollansée et al. (2021), and with meaning recognition scores in a longitudinal study of primary to secondary school learners by De Wilde (2025).

Another out-of-school activity that appears to be positively related to L2 vocabulary knowledge is gaming. Gaming has been found to predict L2 English learners' meaning recognition knowledge (De Wilde et al., 2020b; De Wilde, 2025; Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019; Sundqvist, 2019; Sundqvist & Wikström, 2015), meaning recall knowledge (Kuppens, 2010; Laufer & Vaisman, 2023) and form recall knowledge (Bollansée et al., 2021; González-Fernández & Schmitt, 2015). In terms of vocabulary use, L2 learners who spend more time gaming also tend to produce more varied vocabulary in free writing as measured through TTR (Sundqvist, 2019) and MATTR (Kaatari et al., 2023). Some studies have also indicated that more frequent gamers produced more polysyllabic words in writing (Sundqvist, 2019; Sundqvist & Wikström, 2015), but they did not seem to produce more low-frequency or academic vocabulary (Cobb & Horst, 2011; Olsson & Sylvén, 2015).

It is important to note that research on out-of-school contact has predominantly focused on learners of L2 English. In reviews by Zhang et al. (2021) and Kusyk et al. (2025), it was found that respectively 88% and 66% of the studies included in the reviews had investigated L2 English. Other target languages (e.g., French, Japanese, German) made up only a small portion of the research included in these reviews. This may seem somewhat logical given the status of English as the language of pop culture (Werner & Tegge, 2021) and as 'prestigious lingua franca' (Sundqvist, 2024, p. 6). Yet, this status of English also implies that out-of-school contact with other L2s may be less straightforward or extensive, and the relationship between out-of-school contact and L2 learning may differ for other languages. As Treffers-Daller et al. (2008) state, it cannot be assumed that empirical findings for English can be generalized by default to other L2s. Indeed, two cross-linguistic studies have suggested that out-of-school contact may impact receptive vocabulary knowledge in L2 English and L2 French in different ways (De Wilde et al., 2022; Peters, Noreillie, et al., 2019).

Taken together, Section 2.8. addressed the role of out-of-school L2 contact by defining the construct and discussing research into its relationship with vocabulary knowledge and use. While a positive impact of out-of-school contact has been well attested for meaning recognition knowledge in English, less is known about its impact on productive vocabulary knowledge and vocabulary use. The mixed findings of existing studies indicate that this relationship may be highly complex, and more research is needed to better understand the impact of both overall out-of-school contact and specific out-of-school activities. In addition, the dominant focus on L2 English warrants further research into other L2s, which may hold a different status and may present fewer opportunities for L2 contact.

Conclusion

In conclusion, there is a large body of research into vocabulary knowledge and development in L2 English, with a primary focus on receptive vocabulary knowledge at the level of meaning recognition. There is, however, only limited understanding of the development of productive vocabulary knowledge and vocabulary use in writing, two potentially more challenging levels of mastery that have received less empirical attention. In addition, given the dominance of English in extant L2 research, little is known about the development of vocabulary knowledge and vocabulary use in L2 French. Cross-linguistic studies are even fewer, which complicates the interpretation of findings in light of previous research due to differences in the samples and research designs.

Moreover, in terms of vocabulary use in writing, most studies have examined only one aspect of lexical richness or have only adopted a single approach to examining low-frequency vocabulary use, especially in other L2s than English. More research is needed adopting a multi-dimensional perspective to the analysis of vocabulary use in L2 writing. In addition, few studies to date have focused on the effect of methodological decisions during the pre-processing stage, such as lemmatization and word inclusion, even though they could strongly affect lexical findings and, by extension, the conclusions drawn. In view of the limited research that has directly compared multiple operationalizations of lexical indices in the pre-processing stage, more research is warranted.

3. Methodology

This chapter provides an overview of the methodology employed for the three empirical studies included in the present PhD thesis. The chapter first describes the design (Section 3.1.) and then introduces the research context and participants (Section 3.2.). The following section (Section 3.3.) discusses the data collection instruments in detail. The instruments were selected and adapted based on a pilot study, which is described first, after which the final instruments are presented. Next, Section 3.4. outlines the data collection procedure. The chapter ends with an overview of the lexical analyses (Section 3.5.) by explaining the scoring system for the vocabulary tests and the lexical analyses that were performed on the learners' writing output.

3.1. Design

This PhD project adopted a large-scale, quantitative, correlational design to investigate the relationships between variables through non-experimental methods (i.e., no interventions or manipulations). We adopted a cross-sectional, cross-linguistic, within-subject design. First, for the cross-sectional component, data were collected from learners in two secondary school grades: grade 10 (learners of approximately 15–16 years old) and grade 12 (learners of approximately 17–18 years old). This allowed us to model the differences in productive vocabulary knowledge and vocabulary use between grades, and to increase the generalizability of the findings. Second, the cross-linguistic component involved a focus on two L2s, English and French. This was important in light of the dominant focus on L2 English in extant research (e.g., Plonsky, 2023). Finally, in view of the within-subject design, all data were collected from the same individuals, who each completed a test battery in both L2 English and L2 French. This within-participants approach allowed us to control for confounding variables associated with a between-participants design (e.g., differences in age, language aptitude, working memory).

3.2. Research context and participants

The data collection for the present PhD project took place in Flanders, the northern region of Belgium. In Flanders, Dutch is the official language and the medium of instruction in schools (except in the bilingual Dutch–French region of Brussels, a region which was excluded from the data collection). In Flanders, pupils are taught two mandatory foreign languages in schools, English and French. French is generally introduced in schools from grade 5 (in primary school), around the age of 10 to 11 years old. English instruction starts in grade 7 or 8 (in secondary school) depending on the school, when pupils are approximately 12 to 14 years old.

Throughout secondary school, pupils receive about 3 to 4 weekly hours of French instruction, and about 2 to 3 weekly hours of English instruction in the academic strand. Although Flemish learners thus have longer and more intensive formal French instruction than formal English instruction (in both years and class hours), they tend to be exposed to large amounts of out-of-school English. This seems to be beneficial to receptive vocabulary knowledge even prior to formal instruction (De Wilde et al., 2020a; Peters, Noreillie, et al., 2019; Puimège & Peters, 2019a). Flemish learners are also known to have more positive attitudes towards English than French (e.g., Dewaele, 2005b; Housen et al., 2001), which may contribute to a different L2 learning setting even though both L2s are mandatory subjects in secondary school.

We used convenience sampling to recruit participants, which is a non-probability sampling method that selects participants based on practical criteria like proximity, accessibility and availability (Dörnyei & Csizér, 2012, p. 81). Convenience sampling is the most widely used sampling method in L2 research for reasons of feasibility (DeKeyser et al., 2010; Dörnyei & Csizér, 2012; Plonsky, 2023). We invited school teachers and principals to contribute to the data collection by allowing their class group(s) to participate. Since we were dependent on the willingness and flexibility of the teachers involved, we included any class group whose teacher or principal consented to take part.

Although we used convenience sampling, we took several measures to increase the representativeness and generalizability of the sample. First, we set up a large-scale data collection to ensure a sufficiently large sample size (see below for more detail). Our sample size comfortably exceeds the recommended sample size of at least 100 participants for correlational studies (Durrant et al., 2022). Second, we took care to recruit participants from a wide range of study tracks (including language-related and non-language-related programs), and from schools in varying regions in Flanders (including urban and rural areas).

We originally collected data from 441 participants, of whom 186 were in grade 10 (M_{age} = 15.14) and 255 were in grade 12 (M_{age} = 17.09). However, not all participants were present for the entire data collection. Participants were excluded from the sample if they had not completed all of the tasks in both English and French, or if their writing output in at least one of the L2s consisted of less than 50 tokens. The final sample was thus comprised of 423 L2 learners, of whom 172 were grade 10 students and 251 were grade 12 students. The participants were in 27 different class groups in 9 schools located across Flanders. An overview of the participant characteristics for the full sample can be found in Table 3.1., including

grade, gender, self-reported home languages, study track focus and school region.

For Study 2, which took into account learners' out-of-school L2 contact, the sample size was reduced to 388 participants because some participants indicated that they sometimes spoke English or French at home. Since the use of English or French as a home language cannot be clearly distinguished from out-of-school contact, this was considered a potential concern for the research questions in Study 2. For Study 1 and Study 3, however, these participants were not excluded from the sample for the following reasons. First, the learners took part in general secondary education in Flemish schools with Dutch as a medium of instruction, which means they were considered representative of L2 learners in Flanders, similar to participants with other home languages. Second, for Study 1, we ran the same analyses both including and excluding these participants, and both analyses led to highly similar results. Third, since we asked participants to list the languages they spoke at home with their family (i.e., 'gezin' in Dutch, which typically refers to their parents and siblings), it is unclear to what extent they actually had full conversations in English or French at home. Some learners stated they sometimes integrated English words in their sentences and then classified this as a home language. Since the general findings in Study 1 were not affected by the inclusion of these participants and their inclusion was not deemed a threat to the research questions in Study 1 and Study 3, the participants were only excluded from the sample in Study 2.

For Study 3, which compared different operationalizations of lexical indices, some participants had to be excluded from the sample because their written output was too limited. In Study 1 and Study 2, the full sample of 423 learners had produced at least 50 tokens on the writing tasks in each L2. In Study 3, however, two of the operationalizations excluded function words from the analyses of the learner texts (see Section 3.5.2.). Since this caused some participants' texts to be comprised of fewer than 50 tokens, these participants were removed from the dataset. As a result, the sample size for Study 3 was ultimately reduced to $N = 422$ for English and $N = 398$ for French.

Table 3.1. Characteristics of participants in the full sample ($N = 423$)

	Number of participants
Grade	
Grade 10	172
Grade 12	251
Gender	
Male	163

Female	254
Other / do not wish to disclose	6
Self-reported home languages	
Dutch only	355
Dutch + other	23
Other	10
English and/or French[a]	35
Study track focus[b]	
Language-oriented	135
Grade 10	34
Grade 12	101
Not language-oriented	288
Grade 10	138
Grade 12	150
School region	
City (urban)	186
Municipality (rural)	202

[a] The participants who reported having English and/or French as a home language were excluded from the sample in Study 2 because of its focus on out-of-school contact.
[b] Language-oriented: any study track with a focus on language, i.e., any tracks including Latin, Greek, or modern languages. Not language-oriented: any study track without a focus on language, i.e., science, mathematics, economy, STEM and human sciences.

In addition to the L2 data, we also collected written data from L1 participants as a reference point for vocabulary use (see also Chapter 4). Native speakers of English (N = 30) and French (N = 30) completed the same informal writing tasks as the L2 participants. We collected these L1 data through the online recruitment platforms Prolific (L1 English) and Clickworker (L1 French). Due to the age restrictions of these platforms (viz., accounts can only be created by users over 18 years old), it was not possible to collect data from L1 speakers of the same age as the L2 participants. To ensure that the L1 participants were as comparable as possible to the L2 participants, we strictly adhered to the following inclusion criteria: (1) participants had to be between 18 and 20 years old at the most; and (2) they had to be a student in secondary school or in an undergraduate program. On average, the L1 participants were 19 years old (M_{age} L1 English = 19.17; M_{age} L1 French = 19.33). Since the data were collected via online recruitment platforms, the researcher was not present during test administration. However, participants were asked to complete the task at a time and location where they could not be interrupted or disturbed, and they were explicitly instructed not to leave the online writing environment during the writing task. In order to further motivate the participants to follow the instructions, they received financial compensation for their participation.

3.3. Data collection instruments

3.3.1. Pilot study

Aims of the pilot study

Prior to the large-scale data collection, a pilot study was conducted to assess the suitability of the instruments for our research purposes (Durrant et al., 2022; Malmqvist et al., 2019; Schmitt, 2010b). The aim of the pilot study was three-fold. The first aim was to examine task appropriateness based on learners' performance. The second aim was to verify learners' perception of the tasks, including the perceived difficulty and the learners' motivation for the tasks. The third aim was to collect feedback from a group comparable to the target sample about the clarity of the instructions and the allotted time for each of the tasks.

We piloted the productive vocabulary tests and writing tasks. For the productive vocabulary tests, the goal was to optimize the (existing) tests by eliminating items where necessary (see below) and determining test reliability. For the writing tasks, the goal was to determine which prompts to retain for the final set of writing tasks, as we had not yet made this final selection. For both the vocabulary tests and writing tasks, we used feedback from the pilot participants about the instructions and allotted time to edit the instructions and to set the time limits. The questionnaires used in the final data collection were not piloted, because they have been found to perform well in previously conducted studies (Prophète et al., 2024; Puimège & Peters, 2019a) and they were only minimally adapted for this PhD project.

In total, a convenience sample of 81 participants between 11 and 19 years old took part in the pilot study. Each participant completed one vocabulary test or one writing task in either English (vocabulary n = 8, writing n = 36) or French (vocabulary n = 8, writing n = 29). The vocabulary tests and writing tasks were randomly assigned to the participants.

Productive vocabulary tests

The productive VocabLab tests (Jamers, 2021; Noreillie, 2019; Peters, Velghe, & Van Rompaey, 2019) are lemma-based form recall tests consisting of isolated sentences with a missing word in each sentence (fill-in-the-gap format). They test learners' knowledge at the form recall level of the 3,000 most frequent lemmas in English and French (see Section 3.3.2. about the final instruments for more detailed information about the tests).

Two adaptations were made to the tests prior to the pilot study. First, at the start of this PhD project, the VocabLab test for French was

bilingual and contained L1 translations of the target items, whereas the VocabLab test for English did not contain any L1 translations. In line with Nation's (2013) recommendation to use bilingual tests if possible, and with a view to consistency in the format across English and French, we included an L1 (Dutch) translation of the target word after each sentence (see also e.g., Elgort, 2013; Nguyen & Nation, 2011). In this way, we ensured that participants were not dependent on their L2 reading comprehension skills to be able to provide the correct target form. The L1 translation also minimized the possibility of alternative responses being given. The L1 translations were selected from translations in the online Van Dale translation dictionary (Van Dale, n.d.)[15], taking into account the provided sentence context. The clarity and appropriateness of the translations was verified by asking fellow junior researchers to provide the L2 target word for the chosen L1 translation.

The second adaptation of the original VocabLab tests was related to the form in which the target words were tested. The original tests included some sentences which elicited an inflected form of the target word, for instance, a plural noun (e.g., *des cris de terreur*, *thousands of refugees*), a feminine adjective (e.g., *la plus lente*) or a conjugated verb (e.g., *this seems*, *Olivier a quitté*). However, when target items are morphologically inflected, a vocabulary test can be said to measure more than strictly vocabulary knowledge, as the responses also require morphological knowledge of inflection processes. Since we did not aim to test participants' morphological knowledge of L2 English and French, we adapted these sentences slightly so they would test the base form of the target word (e.g., *a refugee*, *le plus lent*, *Olivier veut quitter*).

Despite the limited sample size (n = 8 per L2), the pilot study served as a general assessment of the suitability of the vocabulary tests for our research purposes. We assessed two aspects of the vocabulary tests: (a) their appropriateness based on an item analysis and test reliability, and (b) learners' feedback on the perceived test difficulty, the test instructions and the time limit.

(a) Item analysis and test reliability

First, we conducted an item analysis in order to examine the reliability of the individual target items based on three criteria: item facility, discrimination index, and Cronbach's alpha if item deleted (Fulcher, 2024; Green, 2019; Noreillie, 2019). The first criterion for the item analysis was item facility, a value between 0 and 1 indicating the proportion of correct answers for an item across participants (e.g., item facility of 0.50 indicates

[15] The Van Dale dictionaries are well-established, renowned and widely used in Dutch-speaking regions like Belgium and the Netherlands.

that half of the participants responded correctly to that item). A low facility value (< 0.20, following Noreillie, 2019) suggests that the item is difficult for the group of test takers; a high facility value (> 0.80) suggests that it is easy. Both situations make it difficult for the item to differentiate between test takers (Green, 2013, 2019). The second criterion for the item analysis was the discrimination index, which reflects the correlation between the score for an individual item and the total test score (Fulcher, 2024, p. 179). A value above $r = 0.30$ is considered good because it means that the item represents the same construct as the overall test (Fulcher, 2024; Noreillie, 2019). The third criterion for the item analysis was Cronbach's alpha if item deleted, a measure of internal consistency when the item in question is removed from the test. If the Cronbach's alpha is higher when the item is deleted than when it is included, this indicates that the item may not be reliable. These three criteria (item facility, discrimination index, Cronbach's alpha if item deleted) were used to determine the suitability of the items on the test for our target sample.

At the start of this PhD, the English and French VocabLab tests were comprised of a different number of target items. The French VocabLab test consisted of 60 target items in total, with 30 items from the 1st-2,000th most frequent lemmas and 30 items from the 2,001st to the 3,000th most frequent lemmas. The number of target items had been reduced from 45 to 30 items per frequency level for Noreillie's (2019) PhD project. Thirty target items are considered sufficiently reliable for testing a 1,000-word frequency band, and this number strikes a balance between reliability and practical considerations such as time constraints (Gyllstad et al., 2021; Nation, 2013; Schmitt et al., 2001). The item analysis based on the pilot study did not reveal any issues with the 60 target items on the French test.

However, the English VocabLab test still contained a total of 90 target items, with 45 items per frequency level. For consistency in the tests across L2s, we used the item analysis to determine which items to remove from the English VocabLab test. We combined the pilot study data ($n = 8$) with data previously collected for a validation study ($n = 62$) by Jamers (2021), in order to eliminate 30 items from the English test (15 items per frequency band). Items that were flagged based on all three criteria (item facility, discrimination index, Cronbach's alpha if item deleted) were removed first, followed by the items that did not perform reliably based on two of the three criteria. Of the items that were flagged by only one criterion, we removed items in order of quality, with the weakest-performing items being removed first, until we reached a total of 30 eliminated items.

In addition to the item analysis to check reliability of individual items, we also verified the reliability of the complete tests. Although

caution is needed due to the small sample size, the internal consistency of the overall tests was high (Cronbach's α English = 0.93; α French = 0.89).

(b) Feedback on test difficulty, instructions and time limit

The second goal of piloting the productive VocabLab tests was to collect participants' feedback on the perceived difficulty of the vocabulary tests[16], on the instructions and on the time limit. In terms of participants' perception of test difficulty, there seemed to be a good balance between participants labeling the tests as easy or as difficult. This was interpreted as an indication that the tests were at an appropriate level of difficulty for the target sample.

The participants indicated that the test instructions were clear and that they all preferred receiving the instructions in their L1 (Dutch) over instructions in L2 English/French. Therefore, no changes were made to the instructions for the vocabulary tests.

In terms of the time limit, participants needed varying amounts of time to complete the vocabulary tests (between 8 and 40 minutes). Since the median time spent on the tests was 19 and 22 minutes for French and English respectively, we retained a maximum time limit of 30 minutes. This allowed faster-working participants to go through the test more quickly, while slower-working participants could take more time if they needed to. Because we anticipated that some participants might spend excessive time on specific items, causing them not to have sufficient time to complete the entire test, we also set a time limit per target item (which was automatized through the online tool used for the data collection). Given that the final tests each contained 60 items with a time limit of 30 minutes, we kept the maximum time identical for all items by setting a 30-second time limit per target item.

Writing tasks

Given that the writing tasks had not yet been selected, the pilot study served as an opportunity to test several potential writing prompts and check their suitability for our research purposes. Since existing research has primarily focused on essay writing, in particular argumentative writing (Heng et al., 2023; Weigle, 2010; Yu et al., 2023), we set out to select prompts eliciting informal writing, a genre that is less well-researched. However, we still piloted a combination of five informal and formal writing prompts in order to compare their suitability for our target sample, based on two criteria: (a) the length of learners' output, and (b)

[16] This refers to the versions of the VocabLab tests at the start of the PhD, namely, the 60-item test for French and the 90-item test for English.

learners' motivation for each of the prompts. Learners' output (text length) was important because the prompts had to elicit sufficient output to allow for automatized lexical analyses (see also Section 3.5.2.). Learners' motivation was also taken into account because higher levels of topic interest and motivation were expected to lead to greater effort and, by extension, more output (Van Steendam et al., 2025).

We drew on previous research to select five informal and formal writing prompts which we believed would be interesting or relatable to the target sample. Two informal prompts were selected from Schoonen et al. (2003; see also Schoonen, 2012), who developed writing prompts in parallel versions to test multiple languages. The selection of prompts for the pilot study was based on rankings of the prompts by adolescents in secondary school and the first year of university. The informal prompts consisted of:

- one prompt in which participants had to write about what they would do if they had won a prize (e.g., a large sum of money in the lottery, or a prize to miss school for a year);
- one prompt in which participants had to complain about something (e.g., their favorite artist not appearing in a famous magazine, or their favorite TV show being canceled).

In addition to these informal prompts, we also selected three formal, argumentative writing prompts inspired by other studies:

- one prompt about social media as a platform to share opinions (Blanchard, 2021);
- one prompt about which factors impact on students' success in education (López-Serrano et al., 2019);
- one self-developed prompt about the quality of remote teaching during the COVID pandemic.

It was important to pilot a range of prompts because we wanted to administer two writing tasks per language for the final data collection, since research has shown that the topic can in itself affect learners' output (e.g., Schoonen, 2012; Yoon, 2017; Zenker & Kyle, 2021). All prompts and instructions were provided in the participants' L1 (Dutch). Each of the writing tasks was completed by five to eight pilot study participants.

The two criteria to select the final writing prompts were the amount of output the participants generated in numbers of tokens (Table 3.2.), and the participants' motivation for the prompts. For English, the formal prompts generally elicited more output than the informal prompts. For French, on the other hand, the output for the formal prompts on success in education and remote teaching was very limited ($Mdn < 100$ tokens). This trend was supported by the feedback provided by the participants, who

repeatedly stated that they thought the topics were difficult and that their French vocabulary knowledge was too limited to write a text about these topics. In addition, multiple participants indicated that they did not find the formal prompts interesting, whereas the informal prompts tended to be appraised more positively (especially the prize prompt). Based on these results, we concluded that the two informal prompts would be the most appropriate choice for our research purposes, because they led to longer texts in French and generally seemed to be more motivating for the participants[17].

Table 3.2. Mean and median number of tokens produced on the writing tasks in the pilot study (n = 65)

	English		French	
	M	***Mdn***	***M***	***Mdn***
Informal prompts				
Prize	210	148	88	117
Complaint	149	115	117	131
Formal prompts				
Social media	256	339	91	122
Success in education	236	169	120	98
Remote teaching	204	167	77	36

Finally, the pilot study also served as a way to check the clarity of the task instructions and the time the participants needed to write their texts. As with the vocabulary tests, most participants preferred receiving the instructions in their L1 (Dutch) because it ensured that they fully and correctly understood the assignment. Another reason to give the instructions in the L1 was to avoid providing any L2 input to the participants, since the goal of the writing tasks was to examine L2 learners' written output without the use of any L2 input or any resources. In terms of the time allotted for the writing tasks, we originally planned to set the time limit for the informal prompts to 20 minutes, following Schoonen (2012). However, the pilot study indicated that this was not always deemed sufficient, since the median time spent on the English informal tasks was between 18 and 20 minutes, and the median time for the French informal tasks was between 22 and 26 minutes. As a result, we set the time limit to 30 minutes per task for the final data collection, in line with the time limit for the vocabulary tests and in accordance with previous research (see e.g.,

[17] We also selected the two formal writing prompts that elicited the most output in English (social media and success in education). Although time constraints did not allow us to incorporate these into the standard data collection procedure, we administered them as English writing tasks to participants on a voluntary basis, if they had finished the other test instruments with sufficient time to spare. The writing output for the formal tasks is beyond the scope of this PhD thesis, but could serve as a basis for a follow-up study.

Fonteyne et al., 2025; Révész et al., 2022). An automatic warning was issued to the participants by the online tool 5 minutes and 1 minute before the time limit.

3.3.2. Final instruments

Productive vocabulary tests

Vocabulary tests were used to measure participants' productive vocabulary knowledge in both English and French. The productive VocabLab tests (Jamers, 2021; Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019) were lemma-based form recall tests of vocabulary breadth. They tested learners' knowledge of the form-meaning link at the level of form recall, which required learners to recall the word form from memory based on a cue (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). The tests were comprised of 60 isolated sentences with one missing word each. The missing words were introduced by their first letter (or two in case of multiple possible responses appropriate to the sentence context), followed by underscores indicating the number of remaining (missing) letters. At the end of each sentence, an L1 (Dutch) translation of the intended target word was provided. Example sentences from the English and French VocabLab tests can be found in Table 3.3.

The productive VocabLab tests were developed in parallel versions for English and French, adhering to the same format and the same item sampling criteria. The tests measured form recall knowledge of target words among the 3,000 most frequent lemmas, based on recent frequency lists by Davies (2008) for English and Lonsdale and Le Bras (2009) for French. From a total of 60 items per test, half of the items (30) were sampled from the 1st to the 2,000th most frequent lemmas, the other half (30) was sampled from the 2,001st to the 3,000th most frequent lemmas. The item selection largely reflected the part-of-speech ratios in the frequency bands, meaning the tests consisted mainly of nouns, followed by verbs and finally adjectives.

Cognates were excluded from the tests to minimize the risk of inflated test scores. Cognates are words that are 'orthographically, phonologically and semantically similar' in two languages (Elgort, 2013, p. 255). On a vocabulary test, this means that an L2 target item resembles a word in the test taker's L1 (e.g., *brilliant* in English, *briljant* in Dutch). It has been shown that cognates can inflate vocabulary test scores (Elgort, 2013; Laufer & McLean, 2016), and they are a potential reason that learners with a limited vocabulary size can still demonstrate knowledge of some low-frequency words (Nguyen & Nation, 2011). While cognates can be argued to be a legitimate part of learners' L2 vocabulary knowledge (e.g., Nation,

2012), they can also distort conclusions about learners' knowledge, because they may cause an overestimation of learners' L2 vocabulary knowledge (e.g., Elgort, 2013; Peters, Velghe, & Van Rompaey, 2019). Moreover, cognates can complicate the comparison of two L2s because (a) languages may contain a different number of cognates with the learners' L1, and (b) the number of cognate items that have been randomly sampled may not be controlled across L2s. For example, Laufer and McLean (2016) found that Japanese learners of L2 English benefitted twice as much from a form recall test with cognates than Hebrew learners.

The full productive VocabLab tests, as administered to the participants, are available in Appendix A1.

Table 3.3. Example sentences from the productive VocabLab tests in English and French

English:	Unfortunately, there is only little research on this very interesting to_ _ _. *(onderwerp)*
French:	Il a découvert une pièce de deux euros dans la p_ _ _ _ de son pantalon. *(zak)* *[He discovered a two-euro coin in the pocket of his trousers.]*

Writing tasks

We selected four informal writing tasks to measure participants' vocabulary use, two tasks in English and two tasks in French (see Table 3.4.). We focused on informal writing because previous research has most commonly investigated argumentative writing, and informal writing has received less empirical attention (Heng et al., 2023; Weigle, 2010; Yu et al., 2023). The pilot study results also indicated that the informal prompts were appropriate for our research sample (see Section 3.3.1.). Another reason for informal writing was that argumentative writing tends to be a highly complex and challenging genre for L2 learners (e.g., Graham & Perin, 2007; Mombaers et al., 2024; Song & Ferretti, 2013), which may limit learners' cognitive resources needed to fully demonstrate their abilities in terms of vocabulary use. Finally, the informal writing topics placed minimal demands on topic knowledge, because learners could draw on their own experiences and interests to choose the angle or focus of their texts.

The prompts were adapted from Schoonen and colleagues (2012; 2003), who developed parallel versions of the prompts to test two languages. These parallel prompts were similar to each other so they would elicit similar output across languages, yet not identical in order to reduce test effects caused by test experience. Specifically, participants were asked

to write about what they would do if they had won a prize (in English, the lottery; in French, a one-year break from school), and to complain about something (in English, their favorite artist being ignored in a magazine; in French, their favorite TV show being canceled). We largely retained the original prompts, since they were already sufficiently concise to limit the linguistic input learners received. This was important given that L2 learners have been found to copy vocabulary from their L1 interlocutor in speaking tasks (Crossley et al., 2016; Lorenzo-Dus & Meara, 2004; Tantucci et al., 2025). By providing limited linguistic input, we avoided unwanted impact of the prompts on the participants' texts (e.g., vocabulary being copied from the prompt or translation effects). While we retained the content of the original prompts, we adjusted the envisioned recipient. In Schoonen (2012), the texts were addressed to formal institutions (e.g., a newspaper, a broadcasting company). Because this could affect the informal nature of learners' texts, we adjusted the recipients to a friend in all four writing tasks.

Table 3.4. Overview of the writing prompts

	Prompts	
English	Prize – version A	Complaint – version A
French	Prize – version B	Complaint – version B

The prompts were presented in the learners' L1, Dutch, to ensure that the participants fully and clearly understood what they were being asked to write about. In addition, since the participants did not receive any L2 input from the instructions, all of the vocabulary in their output can be considered their own, as it was impossible for them to directly copy any vocabulary from the prompts (e.g., Crossley et al., 2016; Lorenzo-Dus, 2007; Tantucci et al., 2025). The writing prompts as they were administered to the participants are available in Appendix A2, accompanied by English translations.

Questionnaires

In addition to the vocabulary tests and writing tasks, we administered a questionnaire for each L2. Since the questionnaires were only minimally adapted from previous studies with learners in Flanders (Prophète et al., 2024; Puimège & Peters, 2019a) after a joint review with the co-authors of the empirical studies, we did not pilot them beforehand. Nevertheless, a researcher was always present during the administration of the questionnaires to answer any potential questions. The questionnaires, similar to the other instruments, were provided completely in Dutch, in order to make sure that participants properly understood all of the

questions. Participants were explicitly encouraged to answer the questions as accurately and honestly as they could.

In total, there were four sections to the questionnaires:

- one section about general language knowledge;
- one section about formal L2 instruction;
- one section about out-of-school L2 contact;
- and one section about attitudes towards the L2.

Firstly, the section about general language knowledge was brief and enquired about languages that learners spoke at home or that they had started learning somewhere (e.g., in school, at a language camp, through Duolingo). We used the self-reported languages spoken at home to determine the learners' home language(s) (see Table 3.1.).

Secondly, in the formal instruction section, we analyzed one question that probed into the onset of formal English and French instruction (learners' first English or French class in school)[18]. The participants' response to this question was used to calculate the number of years of formal instruction they had been exposed to. For example, if a participant in grade 12 indicated that they had had their first English class in grade 8, this was interpreted as 4 years of formal English instruction.

Thirdly, the section about out-of-school L2 contact was comprised of 11 activities for which participants were asked to indicate how much time on average they spent on the activities per week (see Figure 3.1. for an example). For each activity, they had to estimate their weekly time investment in hours and minutes (e.g., 45 minutes, or 2 hours and 30 minutes). We opted for a weekly estimate because selecting a specific day for reference might have led to biased or distorted responses. The weekly estimate also allowed us to account for out-of-school contact on both weekdays and weekends. The L2 activities included in the out-of-school section were based on previous research (e.g., De Wilde et al., 2021; Peters, Noreillie, et al., 2019; Sundqvist, 2019):

- watching TV without subtitles;
- watching TV with L1 subtitles;
- watching TV with L2 captions;
- watching TV with subtitles in another language;
- watching YouTube videos;
- gaming;
- listening to music;

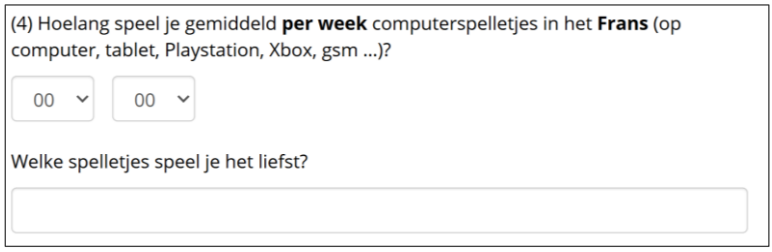
[18] The formal instruction section also comprised questions about what happens in the L2 classroom according to the learners (e.g., how much of the class time the teacher speaks in the L2; which aspects of L2 learning take up the most time during class; how many writing tasks they had to complete in the previous year; and others). These responses have not been analyzed as they were beyond the scope of the present PhD thesis.

- reading;
- listening to or reading social media posts;
- speaking or writing in the L2 to communicate with family or friends (e.g., chatting with a pen pal, calling a long-distance friend);
- speaking or writing in the L2 to create content (e.g., creating a TikTok, writing a story)[19].

Finally, the attitudes section enquired about learners' attitudes towards each of the L2s. This section contained Likert scale questions, for example, about how important, fun, or difficult they think the L2 is, their desire to be good at the L2 and whether they like working for their L2 classes. Given that learner attitudes were beyond the scope of the present PhD project, the responses to the attitudes section have not been analyzed.

The full questionnaire as presented in Dutch can be consulted in Appendix A3, accompanied by a version translated into English.

Figure 3.1. Example item from the section on out-of-school L2 contact on the French questionnaire



*[How much time do you spend on average **per week** playing computer games in **French** (on a computer, tablet, Playstation, Xbox, phone ...)?*
Which games do you prefer to play?]

3.4. Data collection procedure

We set up a large-scale data collection in 27 class groups in schools across Flanders. The participants, L2 learners of English and French in grade 10 (15–16 years old) and grade 12 (17–18 years old), were asked to complete a test battery in both L2 English and L2 French. Data were collected during school hours under the supervision of a researcher. We organized two separate sessions for the full test administration to reduce fatigue effects on the learners' performance. Each session lasted two

[19] In a recent study, Sundqvist and Uztosun (2023) developed and validated the 'Extramural English Scale'. The activities included in our questionnaire seem to align well with the items on the Extramural English Scale.

consecutive class hours (100 minutes) and was dedicated to one L2 (either English or French). The order of the sessions was counterbalanced across class groups to avoid order effects.

The first session started with an introductory presentation by the researcher to give the participants some general information about the research project, with the opportunity for the participants to ask questions. If they agreed to participate, they signed an informed consent form. For students under the age of 16 years old (relevant only in grade 10), the parents also signed an informed consent form which had been sent to them by the teacher beforehand[20]. Students who agreed to participate first took the productive VocabLab test. The order of the target items for each frequency level was randomized per participant. Next, the participants completed two informal writing tasks, whose order was also randomized per participant. Finally, they filled out the questionnaire. The same order of tasks (i.e., first the vocabulary test, then two writing tasks in a randomized order, and finally the questionnaire) was maintained during the second session dedicated to the other L2. The participants were free to take short breaks in between tasks as needed (without interrupting the other participants).

All tasks were completed individually via the online Gorilla Experiment Builder (www.gorilla.sc). The participants had to work individually and without the help of any resources. The researcher was present during test administration to ensure that participants did not consult any (online) tools for help during the tasks. A time limit of 30 minutes was set for the vocabulary test and each of the writing tasks, although participants could work at a faster pace by moving on to the next task whenever they were ready. For the vocabulary test, participants could spend up to 30 seconds on each item. It was not possible to return to a previous item. For the writing tasks, the participants received a warning five minutes before the time limit so they knew their time was running out. In the last minute, an automatic countdown was shown so participants could wrap up their writing in time. There was no time limit for the questionnaires, but most participants were able to complete the questionnaire in approximately ten minutes.

[20] The informed consent procedure and data collection were approved by the Social and Societal Ethics Committee (SMEC) at KU Leuven in March 2022 (approval number G-2020-2902-R3(MIN)). In addition to the informed consent form, students (and where relevant, their parents) also received an information letter with detailed information about how their data would be processed.

3.5. Scoring and lexical analyses

3.5.1. Productive vocabulary knowledge

The productive VocabLab tests were scored dichotomously, with a score of 1 (correct) or 0 (incorrect) per item. On the English vocabulary test, both British-English and American-English variants were approved (e.g., *emphasise* and *emphasize*), even if this implied a difference in the number of letters (e.g., *honor* and *honour*).

We adopted a strict scoring system wherein responses were only awarded a point if they had been spelled correctly, which means only fully known word forms were included in the estimate of participants' productive vocabulary knowledge. Although the scores thus provide a more conservative estimate of learners' productive vocabulary knowledge, the scores reflected a level of mastery at which learners could retrieve the correct orthographic form from memory, which was considered an important aspect of written productive vocabulary knowledge (e.g., Johnson et al., 2016; Nation, 2013; Webb, 2008). In a comparison of a sensitive and a strict scoring system, Webb (2008) found that a strict scoring system seemed to better reflect differences in productive form recall knowledge between two groups with the highest and lowest receptive vocabulary test score, respectively. Moreover, while strict scoring is not sensitive to partial knowledge, it is straightforward in that responses are either entirely correct or deemed incorrect. This means that raters do not have to make a subjective judgement of which errors should be awarded a point[21]. In addition, a study by Johnson et al. (2016) found that holistic scores of writing performance were more strongly related to vocabulary test scores based on a strict scoring system than on a more lenient system. This suggests that it is useful to operationalize vocabulary knowledge as full *accurate* knowledge of meaning and form. The total test score out of 60 items was converted to a percentage score to facilitate interpretation. The same test scores were used for all three empirical studies included in this PhD thesis.

[21] A lenient scoring system requires raters to determine which errors are deemed acceptable to be awarded a point (e.g., 'a spelling error which does not distort the word'; Laufer, 1998, p. 260). This also applies to other constructed-response tests like meaning recall, where researchers have to decide what counts as a demonstration of knowledge of meaning (e.g., for the word *screwdriver*, responses could include *a device to turn screws, a tool, something to fix things,* or even *a drink*). As a result, it is crucial to have clear rater instructions and to report inter-rater reliability if any subjectivity is inherent to the scoring system (e.g., Durrant et al., 2022).

3.5.2. Vocabulary use

Pre-processing of learner texts

Although the participants completed two informal writing tasks per L2, the two texts were combined into one file and analyzed as a single text. The first reason for this was to keep the design and the statistical models as parsimonious as possible. A second reason was to limit the data loss as a result of the limited output provided by some learners. We set the lower limit of text length to 50 tokens, because this was the minimum number of tokens required to calculate lexical diversity (see below). Since a number of participants failed to reach a total of 50 tokens even when the two tasks were taken together (in L2 English or French, sometimes in both), there would have been considerably more data loss if the two tasks had been analyzed separately.

In line with the strict scoring system for the vocabulary tests, the learners' writing output was analyzed exactly as it had been submitted by the participants (albeit in an anonymized version) (Woods et al., 2023). Spelling errors were not corrected, because we aimed to examine learners' written vocabulary use exactly as they had included it in their output and thus, how they believed the words were supposed to be written and used. Although the use of incorrectly spelled words may reflect partial knowledge, it also suggests that learners do not have 'full knowledge of meaning and form' of those words (Webb, 2008, p. 85). It can therefore be argued that such words should not be included in their corrected form in an analysis of learners' vocabulary use.

This strict approach also avoided any subjectivity in the decision on which errors to correct. For instance, Laufer and Nation (1995) omitted words in the learner texts that were used in a semantically inappropriate context because these words 'could not be considered as part of the subject's productive lexicon' (p. 135), whereas they did correct spelling errors, implying that misspelled words were interpreted as known by the participants (the authors refer to this as 'familiar' rather than 'known'). It seems rather arbitrary to apply sensitive scoring to matters of orthographic accuracy (i.e., correcting spelling errors, related to form), but strict scoring to matters of semantic accuracy (i.e., omitting semantically incorrect words, related to meaning), because knowledge of both form and meaning are inherently part of being able to use vocabulary in writing (Nation, 2013). Through a strict scoring system, a more neutral and consistent approach can be taken to a lack of (partial) orthographic or semantic knowledge by not correcting either type of error in the learners' texts.

We used the lemma as the counting unit for all of the lexical analyses, in accordance with the lemma-based VocabLab tests. Lemmas are

word clusters comprised of the base form of a word and its inflections in the same part-of-speech. Lemma-based tests and analyses have been gaining traction in recent years (e.g., Kyle et al., 2018; Peters, Velghe, & Van Rompaey, 2019), possibly due to an increasing awareness (or even, problematization) of the implicit assumption of word families that learners have full (productive) knowledge of the L2 inflectional and derivational system (e.g., Iwaizumi & Webb, 2022; Kyle, 2020b; McLean, 2018; Schmitt & Zimmerman, 2002).

The learner texts were tokenized, part-of-speech tagged and lemmatized using the spaCy models en_core_web_lg (English) and fr_core_news_lg (French) (Honnibal et al., 2020). SpaCy is an open-source library that can be used for natural language processing in the programming language Python[22]. The tokenization process consists of spaCy dividing a given text into individual words or so-called tokens (see Table 3.5. for example sentences). During part-of-speech (POS) tagging, spaCy determines each word's POS based on the sentence context (e.g., *seat* can be used as a noun or verb, *fast* can occur as an adjective or an adverb). Finally, lemmatization entails categorizing each token under a headword based on its raw form and the POS. For example, the raw token *runners* would be categorized under the nominal lemma *runner*; the token *ran* would be categorized under the verbal lemma *run*. The spaCy models used both have high tagging accuracy, with the spaCy website reporting 100% tokenization accuracy and above 97% POS tagging accuracy[23]. We used the default lemmatization integrated in spaCy for all three empirical studies included in this PhD thesis.

Table 3.5. Example sentence demonstrating tokenization and POS tagging in spaCy

	English	**French**
Raw sentence	*I'm training to run faster.*	*Je m'entraîne pour courir plus vite.*
Tokenized	I	Je
	'm	m'
	training	entraîne
	to	pour
	run	courir
	faster	plus

[22] We used ChatGPT (OpenAI, 2024) to help adapt the Python programming codes.

[23] All accuracy rates can be consulted from the spaCy website via https://spacy.io/models/en and https://spacy.io/models/fr. The transformer models (TRF) have even higher accuracy rates, but they require more processing power, especially for larger-scale analyses. For a more detailed accuracy analysis of the English models, see Kyle and Eguchi (2024).

		vite
POS tagged and lemmatized	I_PRON	je_PRON
	be_AUX	me_PRON
	train_VERB	entraîner_VERB
	to_PART	pour_ADP
	run_VERB	courir_VERB
	fast_ADV	plus_ADV
		vite_ADV

In Study 3, we compared two lemmatization schemes, a default and custom lemmatization scheme. Under default lemmatization, lemmas were distinguished following the standard definition of a lemma, that is, the base form and its inflections in the same POS. The custom lemmatization scheme distinguished verbal lemmas not only based on the base form and POS, but also based on their mood and tense (see Table 3.6. for examples in English and French). This lemmatization scheme accounted for the complex verbal inflectional system in French especially (see Consolini & Kyle, 2024, for a similar approach with L2 Spanish). For instance, while the conjugated forms of *see* (e.g., *see, sees, seeing, saw, seen*) would all be classified under the same headword under default lemmatization, the custom lemmatization scheme distinguished *see* in the infinitive, *see* and *sees* in the indicative present, *seeing* as present participle, *seen* as past participle, and *saw* in the indicative past. This was especially relevant for French, which is a morphologically rich language with a complex inflectional system. The default lemma for *voir* in French is comprised of over twenty inflected forms in different moods and tenses (e.g., *vois*, *voit, voie, voyais, verrai, verrais, voyant, vu*, etc.). The custom lemmas, on the other hand, distinguish between these moods and tenses (e.g., separate custom lemmas for conjugated forms in the indicative present, the indicative past and the subjunctive present). The custom lemmatization scheme represents a lower-level counting unit and thus accounts to some extent for the rich verbal morphology in French.

Table 3.6. Illustration of default and custom lemmatization in Study 3

Default lemmatization		**Custom lemmatization**	
English: *to see*			
see_VERB:	see, sees, seeing, seen, saw	see_VERB_Inf:	see
		see_VERB_Ind_Pres:	see, sees
		see_VERB_Ind_Past:	saw
		see_VERB_Part_Pres:	seeing
		see_VERB_Part_Past:	seen

French: *voir*			
voir_VERB:	voir, vois, voit, voyons, voyez, voient, voyais, voyait, voyions, voyiez, voyaient, voyant, vu, verrai, verras, verra, verrons, verrez, verront, verrais, verrait, verrions, verriez, verraient, voie, voies, voyions, voyiez, voient, etc.	voir_VERB_Inf:	voir
		voir_VERB_Ind_Pres:	vois, voit, voyons, voyez, voient
		voir_VERB_Ind_Past:	voyais, voyait, voyions, voyiez, voyaient
		voir_VERB_Part_Pres:	voyant
		voir_VERB_Part_Past:	vu
		voir_VERB_Ind_Fut:	verrai, verras, verra, verrons, verrez, verront
		voir_VERB_Ind_Con:	verrais, verrait, verrions, verriez, verraient
		voir_VERB_Sub_Pres:	voie, voies, voyions, voyiez, voient

Apart from lemmatization, a second consideration during the pre-processing stage which was investigated in Study 3, was which words to include in the analyses. By default, all words in the learner texts were included. In Study 3, we also compared lexical findings based on word inclusion, whereby we performed the lexical analyses either on all words or on content words only. All learner texts were thus analyzed with all words included, and with only the content words included. In line with existing research, we defined content words as nouns, lexical verbs (no auxiliaries), adjectives and adverbs (Engber, 1995; Kyle, 2020b; Kyle & Eguchi, 2023). Based on spaCy's POS tags, the function words (i.e., all words not tagged by spaCy as 'NOUN', 'VERB', 'ADJ', or 'ADV') were excluded from the content word analysis.

Lexical diversity

The first construct that was used to measure learners' written vocabulary use was lexical diversity. Lexical diversity is the variety of words or the level of lexical repetition in a text (though see Jarvis, 2013, 2017, for a broader interpretation). We analyzed lexical diversity using the moving-average type-token ratio (MATTR). MATTR is a variation of the original

type-token ratio that has been shown to be highly stable across text lengths (e.g., Bestgen, 2024; Kyle et al., 2021; Zenker & Kyle, 2021), thereby countering a widely acknowledged limitation of the original type-token ratio. MATTR is the average of type-token ratios of contiguous text segments of equal length (commonly 50 words for written texts). This means that type-token ratios are calculated for words 1–50, then words 2–51, words 3–52, words 4–53 until all words have been included in a calculation (for a text of 100 words, that would mean the last segment consists of words 51–100). The MATTR value is the average of all of these individual type-token ratios. Table 3.7. shows an example of the MATTR calculation (based on 5-word segments for illustration purposes). MATTR was calculated on the lemmatized learner texts using Kyle et al.'s (2021) lexical diversity Python package.

Table 3.7. Example of MATTR calculation based on lemmatized 5-word segments

	The dog chased the cat, but the cat was faster than the dog.		
	Lemmatized segment	**Types/tokens**	**TTR**
Tokens 1–5	the dog chase the cat	4/5	0.80
Tokens 2–6	dog chase the cat but	5/5	1
Tokens 3–7	chase the cat but the	4/5	0.80
Tokens 4–8	the cat but the cat	3/5	0.60
Tokens 5–9	cat but the cat be	4/5	0.80
Tokens 6–10	but the cat be fast	5/5	1
Tokens 7–11	the cat be fast than	5/5	1
Tokens 8–12	cat be fast than the	5/5	1
Tokens 9–13	be fast than the dog	5/5	1
		MATTR = 0.89	

A disadvantage of analyzing vocabulary use only in terms of lexical diversity, is that lexical diversity indices such as MATTR treat all words in a text as equally difficult (Kyle, 2020b). More concretely, MATTR is not able to distinguish between the sentence *The dog chased the cat, but the cat was faster than the dog* and the sentence *The hound pursued the feline, yet the feline proved swifter than the hound* (both 13 tokens with 8 types; both a MATTR value of 0.89). While both sentences are identical in terms of lexical diversity, it is more likely that the second sentence would be written by a more proficient L2 learner than the first sentence. Scholars have therefore called for a multivariate approach to measuring vocabulary use, adopting indices of both lexical diversity and lexical sophistication for a more comprehensive picture of learners' vocabulary use (e.g., Consolini & Kyle, 2024; Jarvis, 2013; Kyle, 2020b).

Lexical sophistication

The second construct that was used to measure learners' vocabulary use was lexical sophistication. Indices of lexical sophistication are a useful addition to indices of lexical diversity because they account for the different types of words that are used (e.g., *dog* versus *hound*). Lexical sophistication is the use of sophisticated or advanced words, most commonly interpreted as the use of low-frequent words (though see Kyle, 2020b, for an overview of other word properties related to sophistication). The frequency-based interpretation of sophistication has previously been operationalized in two ways: through a band-based approach and through a mean-based approach.

The band-based approach is generally referred to as Lexical Frequency Profile or LFP (Laufer & Nation, 1995). The LFP draws on frequency bands that are created based on frequency counts in a reference corpus. A word's frequency of occurrence in a reference corpus determines its position in a list of words ordered by decreasing frequency. For instance, a word appearing 200,000 times in the reference corpus will be placed closer to the top of the ranked frequency list than a word with 2,000 occurrences. Each word in the frequency list is then assigned a frequency band: the first 1,000 most frequent words make up the first frequency band (1K), the following 1,000 words the second band (2K), the next 1,000 words the third band (3K), and so forth. The LFP provides an overview of the proportion of words in a learner's text appearing in each of the frequency bands (e.g., 85% of words in 1K, 10% of words in 2K, 3% in 3K, 1% in 4K, 1% in 6K and 9K). To facilitate statistical models and interpretation of LFP results, we used a condensed LFP with one value representing the proportion of words beyond the 2K level (Laufer, 1995, 1998). A higher LFP value should be interpreted as a higher level of lexical sophistication.

The band-based approach to lexical sophistication has received some criticism for being coarse-grained, because it is based on 1,000-word bands (Crossley et al., 2013; Meara, 2005a). Researchers are now increasingly using a mean-based approach to lexical sophistication, by calculating the average frequency of the words in a learner's text. This approach uses direct frequency counts from a reference corpus rather than calculating the proportion of words beyond a certain frequency band. Each word in the learner's text is checked for its frequency of occurrence in the reference corpus, and the average frequency for all words in the text is then calculated. In contrast with the LFP, a higher average frequency should be interpreted as a lower level of lexical sophistication (higher frequency values indicate that words are more common in the reference corpus and, thus, less sophisticated). Mean-based measures may be better at distinguishing between proficiency levels, although Crossley et al. (2013)

note that they also provide less apparent pedagogical applications. Crossley et al. therefore recommend taking a multivariate approach, by adopting both band-based and mean-based measures. Another reason to combine the LFP and the average frequency in this PhD thesis was to compare to what extent the results for a band-based and mean-based measure might differ when they are computed on the same learner texts.

While many studies have analyzed frequency-based lexical sophistication in L2 English writing, fewer studies have analyzed word frequency in other L2s such as French. This is also reflected in the limited number of tools available to analyze lexical sophistication in L2 French, although some efforts have been made to create lexical analysis tools for other L2s than English. Two examples are Cobb's (n.d.) Compleat Web VocabProfile and Finlayson et al.'s (2023) MultilingProfiler[24]. The online Compleat Web VocabProfile offers frequency lists in both English and French. However, Cobb is known for continuously updating his website and tools, making it difficult to outline the exact underlying processes during analyses, which complicates replication of findings. In addition, in terms of lemma-based frequency lists available from the website, the reference corpora are limited to the British National Corpus for English and Lonsdale and Le Bras's (2009) corpus for French. While these are widely used corpora, some research has shown that frequency lists based on subtitle corpora are more strongly related to L2 learner knowledge, and that they more accurately predict reaction times in lexical decision tasks (e.g., Brysbaert & New, 2009; Pinchbeck et al., 2022). The second tool, MultilingProfiler, a tool designed for the analysis of French, German and Spanish, partly accounts for this by allowing users to upload their own custom frequency lists to the website. Nevertheless, because MultilingProfiler is not available for English, the use of this tool would have compromised the consistency in the procedure across the English and French analysis. A final practical consideration was that the aforementioned tools are both mainly targeted at language teachers or smaller-scale analyses, so they do not allow for batch processing. As a result of these limitations, we developed our own natural language processing pipeline in Python, which allowed us to fully control and customize features like the reference corpus and the counting unit, and to batch process the full learner corpus at once.

We used two reference corpora to calculate word frequency counts: Subtlex-US (Brysbaert & New, 2009) and Lexique (version 3.83) (New et al.,

[24] The Compleat Web VocabProfile is available from https://www.lextutor.ca/vp/comp/. MultilingProfiler is available from https://www.multilingprofiler.net/. In July 2025, a new tool LexPro was published by Van Parys et al. (2025) allowing flemma-based analyses of English, French, Dutch and Spanish. LexPro uses a band-based approach to analyze frequency-based lexical sophistication.

2007). These corpora are comprised of subtitles from a wide range of TV series and films, and they are made up of 51 and 52 million tokens respectively. The frequency counts for these corpora are readily available online, although frequencies based on lemmas (rather than word forms) could only be consulted for Lexique, not Subtlex-US. For Study 1 and Study 2, we used the lemma frequencies (normalized per million words) available online for Lexique, and we calculated the lemma frequencies (per million words) for Subtlex-US ourselves using the default spaCy settings. We calculated the average log frequency of each learner text by averaging the log-transformed lemma frequencies of the words appearing in the text. The log-transformed values were used rather than raw values to account for skewness typical of frequency counts (Zipf's law). In order to calculate the LFP, we ranked the lemma frequency lists in order of decreasing frequency and assigned a frequency band to each lemma (grouped per 1,000 words). We distinguished 1,000-word bands for frequency levels up to 15K and created one larger band for words beyond 15K. The condensed LFP was the summation of the proportion of words in each frequency band beyond 2K (i.e., from 3K to 16K+). Off-list words were excluded from the LFP calculation.

In Study 3, we tweaked the Python pipeline based on Kyle's (2025) Pylats Python package (e.g., we defined which spaCy tags should be considered content words, we excluded non-existent words tagged as 'X' by spaCy, and ignored proper nouns for the analyses). We analyzed the reference corpora and the learner corpus using the exact same processing pipeline. We first tokenized, POS tagged and lemmatized the reference corpora (Subtlex-US and Lexique) using spaCy models, which allowed us to create word lists containing the lemma frequency counts (normalized per million words). Lemmatizing the reference corpora ourselves was especially important for the customization of the lemmatization scheme, since the default and custom lemmatization resulted in different lemma frequency lists. For instance, a lemma under default lemmatization would be *see_VERB* (only headword and POS), while this would be *see_VERB_Ind_Pres* under custom lemmatization (headword, POS, and mood/tense information). The same lemmatization scheme had to be used for both the reference corpora and the learner corpus to ensure reliability of the frequency counts. Another advantage of using the same processing pipeline for the reference corpora as for the learner texts is that it ensures consistency in how lemmas are defined, for example, in the case of contracted forms (e.g., whether *I've* is lemmatized as *I have* or *I 've*, which can impact on the frequency counts). The resulting frequency lists were used to calculate the average log frequency of the learner texts (average of the log-transformed lemma frequencies), and to compute the condensed LFP (proportion of lemmas beyond 2K). Off-list words were excluded from the final calculations.

Summary

In summary, the present PhD project presents a large-scale and multi-dimensional investigation of L2 learners' productive vocabulary knowledge and vocabulary use, in a cross-sectional, cross-linguistic within-subject design. One overarching large-scale data collection was organized, in which a sample of 423 participants from grade 10 and grade 12 completed a productive vocabulary test, two informal writing tasks and a questionnaire in both L2 English and L2 French. Productive vocabulary knowledge was measured at the level of form recall and was represented by a total percentage score on the vocabulary test. Vocabulary use in writing was measured in terms of lexical diversity (moving-average type-token ratio) and two indices of frequency-based lexical sophistication (average log frequency, lexical frequency profile). The following chapters report on the three empirical studies that were conducted as part of this PhD project, based on this data collection.

4. Vocabulary knowledge and vocabulary use in writing: A cross-sectional study[25]

Abstract

This study investigated the lexical proficiency of L2 learners of English and French. The aim of the study was two-fold. First, we examined the cross-sectional differences in productive vocabulary knowledge and vocabulary use between L2 learners in two grades. Second, we investigated the extent to which vocabulary knowledge and grade could predict vocabulary use in writing, operationalized as lexical diversity (moving-average type-token ratio), mean-based and band-based lexical sophistication (average frequency and lexical frequency profile). Participants (N = 423) from grade 10 and grade 12 completed a form recall test (productive vocabulary knowledge) and two writing tasks (vocabulary use) in both English and French. The cross-sectional comparison showed that while vocabulary knowledge and lexical diversity increased across grades in both languages, mean-based lexical sophistication only improved in English. Band-based sophistication was not observed to increase in either language. Further, the results indicated that productive vocabulary knowledge predicted all measures of vocabulary use in English, but only predicted lexical diversity for French. The findings are discussed in light of their implications for L2 pedagogy.

[25] Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., Puimège, E., Creten, S., & Peters, E. (2025). Vocabulary knowledge and vocabulary use in writing: A cross-sectional comparison of L2 English and L2 French. *Applied Linguistics, Advance online publication,* amaf009. https://doi.org/10.1093/applin/amaf009

4.1. Introduction

Vocabulary is widely recognized as a key aspect of second language (L2) acquisition and has become an important topic of investigation. Research has suggested that vocabulary learning is incremental, and different components of word knowledge may not develop at the same rate (e.g., González-Fernández & Schmitt, 2020). There is some evidence from cross-sectional studies that productive vocabulary knowledge (i.e. form recall knowledge) and vocabulary use (i.e. using words in one's own output) develop at a different pace. Two cross-sectional studies have compared productive vocabulary knowledge and vocabulary use in writing by testing learner groups in different grades of secondary school (Henriksen & Danelund, 2015; Laufer, 1998). However, their small sample sizes impede strong conclusions about the generalizability of the findings.

It stands to reason that vocabulary knowledge is a prerequisite for L2 writing. Multiple studies have shown that receptive vocabulary knowledge is a strong predictor of L2 writing performance (e.g., Stæhr, 2008; Miralpeix & Muñoz, 2018). However, much less is known about the role of L2 learners' productive vocabulary knowledge in L2 writing, although it could be particularly important in relation to L2 learners' written vocabulary use. With the exception of a few studies that have returned conflicting findings (Laufer, 1998; Henriksen & Danelund, 2015; Karakoç & Köse, 2017), there is little research investigating the relationship between productive vocabulary knowledge and vocabulary use. Moreover, previous studies on vocabulary knowledge and use are strongly geared towards L2 English. Yet, there is a need for cross-linguistic studies to examine whether the findings can be generalized across L2s.

The present study aims to contribute to our understanding of lexical proficiency by comparing L2 learners in two grades in terms of productive vocabulary knowledge and vocabulary use of single words. Additionally, the study adds to existing research by exploring the relationship between L2 vocabulary knowledge and vocabulary use in writing in two L2s: English and French. As the first large-scale, cross-sectional and cross-linguistic study on L2 vocabulary knowledge and vocabulary use, this study will help to shed light on productive vocabulary knowledge and vocabulary use in writing in two L2s by learners at different stages of L2 instruction.

4.2. Literature review

4.2.1. L2 vocabulary knowledge

In this study, we define a second language (L2) as any additional language learned after the first language (L1), regardless of other known languages and the learning context (for a more detailed account of L2

definitions, see Ortega, 2013). L2 vocabulary learning is often regarded as an incremental process in which vocabulary knowledge develops from receptive to productive knowledge, following Nation's (2013) framework. Receptive vocabulary knowledge refers to the ability to understand a word when it is presented in spoken or written form, and is usually measured by tests of meaning recognition (choosing the meaning from multiple options) or meaning recall (describing the meaning). Productive vocabulary knowledge refers to learners' ability to produce a word themselves and is generally tested as form recall, where learners have to retrieve a word from memory based on a stimulus (e.g., a sentence or description). Research has repeatedly shown that receptive knowledge precedes productive knowledge, as meaning recognition and meaning recall knowledge develop before form recall knowledge (e.g., González-Fernández & Schmitt, 2020; Webb, 2008).

There is some empirical research into L2 learners' development of vocabulary knowledge (see also Miralpeix, 2020). In a five-year longitudinal study, Webb and Chang (2012) demonstrated that L2 English learners' meaning recognition scores improved yearly throughout secondary school and university. In a cross-sectional design, Peters, Noreillie, et al. (2019) showed that university students obtained higher meaning recognition scores in L2 English and French than grade 10 secondary school students, who in turn outperformed grade 8 learners. Although research on productive vocabulary knowledge is limited, Laufer (1998) found that grade 11 learners scored higher than grade 10 learners on a meaning recognition and form recall test. Further, Laufer's study indicated that recognition and recall may develop at different rates. Recall scores were generally lower than recognition scores and also showed a smaller cross-sectional difference than recognition scores, suggesting that recall knowledge develops less quickly than recognition knowledge. However, due to the lack of cross-sectional studies investigating productive vocabulary knowledge, in L2 English but also in other L2s, more research is necessary.

4.2.2. L2 vocabulary use

Vocabulary *use* is another component of word knowledge (Nation, 2013), defined as a learner's ability to produce words in their own speech or writing. This study only considers use of single words, although there is a growing body of work on the use of multi-word units (see e.g., Bestgen & Granger, 2014; Siyanova-Chanturia, 2015). Vocabulary use is often operationalized as lexical richness, which can be divided into three constructs: density, diversity and sophistication (Read, 2000). First, lexical density, the proportion of content words in a text, has not consistently been

shown to be related to L2 proficiency and may thus not be ideal as a measure of L2 writing performance (Kyle, 2020b). Second, lexical diversity reflects the variety of words in a text, often operationalized as the type-token ratio, which measures the proportion of unique words (types) against the total number of words (tokens). As the original type-token ratio has been criticized for its dependence on text length, several alternative indices have been developed (e.g., Guiraud's Index, HD-D, MATTR, MTLD). The moving-average type-token ratio (MATTR) and measure of textual lexical diversity (MTLD) have been shown to be particularly stable across text lengths (Bestgen, 2024; Kyle et al., 2021; McCarthy & Jarvis, 2010; Treffers-Daller, 2013; Zenker & Kyle, 2021). Finally, lexical sophistication refers to the use of sophisticated words, often measured as low-frequency words. Words in a learner text occurring less frequently in a reference corpus (e.g., British National Corpus) are considered more sophisticated and thus more advanced.

Both diversity and sophistication seem to be related to L2 proficiency. Research has demonstrated that lexical diversity in learners' writing output is positively linked to L2 proficiency (e.g., Consolini & Kyle, 2024; Treffers-Daller et al., 2018; Zenker & Kyle, 2021) and holistic ratings of L2 writing (e.g., Daller & Phelan, 2007; Engber, 1995; Guo et al., 2013; Treffers-Daller et al., 2018). Likewise, a positive relationship has been established between lexical sophistication and holistic scores of L2 writing (e.g., Crossley & McNamara, 2012; Guo et al., 2013; Johnson et al., 2016; McNamara et al., 2010). Additionally, a study by Lindqvist (2010) showed that lexical sophistication in L2 French could distinguish between L2 proficiency levels. However, in contrast with Lindqvist (2010), two small-scale cross-sectional studies have indicated that lexical sophistication in writing did not improve significantly between grades (Henriksen & Danelund, 2015; Laufer, 1998). Given the lack of large-scale analyses of vocabulary use, it remains to be seen whether these findings would be corroborated in a study with a larger sample.

Moreover, many of the aforementioned studies used band-based measures of lexical sophistication, as they determined the lexical frequency profile (LFP) of learners' texts (Laufer & Nation, 1995). LFP calculates the proportion of words in a learner text belonging to 1,000-word frequency bands based on a reference corpus. An alternative to the band-based LFP is a mean-based approach, whereby the average frequency of learners' texts is calculated based on frequency counts in a reference corpus (i.e. a measure based on counts rather than bands). As count-based indices may be better at distinguishing between proficiency levels than band-based indices (Crossley et al., 2013), different findings may emerge based on a mean-based measure of sophistication.

Further, most studies on vocabulary use in writing so far have focused on one L2 only (usually English), which raises the question whether the results can be extended to other L2s. While a few studies have examined vocabulary use in spoken and written L2 French or L2 Spanish (Consolini & Kyle, 2024; Lindqvist, 2010; Noreillie et al., 2020; Tracy-Ventura, 2017), there have been no cross-linguistic studies integrating multiple L2s. Moreover, conclusions from research on L2 vocabulary use have mostly been based on data from participants who had been asked to complete a single writing task. Research has indicated that more familiar topics may elicit more advanced vocabulary than unfamiliar topics, and certain topics tend to yield more lexically diverse texts than others (Yoon, 2017; Zenker & Kyle, 2021). Yet, the impact of topic on learners' output is often overlooked in L2 vocabulary research. Given the limited attention to these factors, it is uncertain whether the findings from previous research can be generalized across writing topics and L2s.

4.2.3. Vocabulary knowledge and use

Research indicates that vocabulary knowledge plays an important role in L2 proficiency. Receptive vocabulary knowledge has been evidenced to be a strong predictor of L2 language skills (Stæhr, 2008; Miralpeix & Muñoz, 2018; Lin, 2023; Noreillie et al., 2020). Stæhr (2008) showed that meaning recognition scores were moderately associated with listening skills and strongly associated with reading and writing performance. Findings by Miralpeix and Muñoz (2018) confirmed that receptive vocabulary knowledge was a predictor of L2 performance in listening, reading, speaking and writing. The relationship was strongest for writing performance, which may indicate that receptive vocabulary knowledge is especially crucial to L2 writing development. Similarly, a few studies have demonstrated the importance of productive vocabulary knowledge in L2 writing performance (e.g., Johnson et al., 2016; Karakoç & Köse, 2017).

Only a limited number of studies have specifically investigated the relationship between vocabulary knowledge and vocabulary use. For L2 spoken output, a study by Uchihara and Clenton (2020) found no correlation between receptive vocabulary knowledge and lexical sophistication in L2 spoken output, whereas Uchihara et al. (2020) found a correlation between productive vocabulary knowledge and both lexical diversity and sophistication. In terms of L2 writing, Laufer and Nation (1995) was one of the first studies to investigate the relationship between vocabulary knowledge and vocabulary use. L2 learners of English completed a productive vocabulary test and two writing tasks. Vocabulary use was operationalized as lexical sophistication by means of the lexical frequency profile using word family lists (LFP; band-based measure).

Laufer and Nation's (1995) study demonstrated a positive correlation between productive vocabulary knowledge and lexical sophistication, as learners with more vocabulary knowledge produced fewer high-frequent words and more low-frequent words.

In a later study, Laufer (1998) administered a receptive and productive vocabulary test and one writing task to L2 English participants. Laufer (1998) used a condensed LFP (band-based) with one value representing the proportion of low-frequency words, and analyzed the proportion of words 'beyond 2K' (beyond the 2,000 most frequent word families). Although this operationalization entails some loss of information compared to the full LFP, an important advantage of the condensed LFP is that it facilitates statistical analyses and interpretation of the findings. Laufer's study did not reveal a correlation between learners' vocabulary knowledge and lexical sophistication in writing. This may be partly explained by the sample size (N = 48), which may have been too small to detect any relationship. In contrast, one other small-scale study did demonstrate a positive relationship between productive vocabulary knowledge and band-based lexical sophistication (Henriksen & Danelund, 2015). Additionally, Henriksen and Danelund (2015) found that productive vocabulary knowledge was positively related to lexical diversity, measured as Guiraud's Index to account for text length differences. However, Guiraud's Index may still be dependent on text length because it is calculated based on the text as a whole, rather than contiguous same-size text segments (Bestgen, 2024). In sum, while productive vocabulary knowledge may be an important factor in lexical diversity and sophistication in L2 learner output, previous studies have not always provided conclusive empirical evidence of this relationship.

4.3. Research questions

The present study aimed to gain insight into the lexical proficiency of L2 learners of English and French by examining their productive vocabulary knowledge and vocabulary use in writing, with a focus on single words. Vocabulary knowledge was measured at the form recall level. Vocabulary use was operationalized as both lexical diversity and lexical sophistication. The following research questions were addressed.

RQ1. What are the cross-sectional differences between grade 10 and grade 12 learners in terms of productive vocabulary knowledge, lexical sophistication and lexical diversity in written L2 English and L2 French?

RQ2. To what extent do productive vocabulary knowledge and grade predict lexical sophistication and lexical diversity in written L2 English and L2 French?

4.4. Method

4.4.1. Design and procedure

Data were collected in two grades to allow a cross-sectional comparison of two learner groups who differ in length of instruction (see Participants). In order to increase generalizability across L2s, we adopted a within-participants design in which participants completed all tasks in both English and French. The within-subject design also allowed us to control for confounding variables associated with a between-participants design (e.g., language aptitude, working memory, age).

Data collection took place in class groups during school hours under supervision of a researcher. To reduce fatigue effects on participants' performance, we organized two data collection sessions of two consecutive class hours (100 minutes per session), with approximately one week between the sessions. Each session was dedicated to either English or French. The language order was counterbalanced to minimize order effects. All tests were completed on a laptop or tablet via the online platform Gorilla Experiment Builder (www.gorilla.sc). During each session, participants first completed a form recall vocabulary test, followed by two informal writing tasks in the same language (i.e. two topics per language; see Instruments). The order of the writing tasks was randomized via the online platform.

4.4.2. Participants

Participants were recruited from secondary schools in Flanders, the Dutch-speaking region of Belgium, where Dutch is the official language and medium of instruction in schools.

We collected data from a convenience sample (N = 441) consisting of 186 grade 10 students (M_{age} = 15.14) and 255 grade 12 students (M_{age} = 17.09). Participants were excluded from the present study if they had not fully completed both sessions (i.e. all tasks for both languages), or if they had produced fewer than 50 tokens on the writing tasks in at least one language. This resulted in a full sample size of 423 learners.

The participants in grade 10 (15–16 years old) had had 5 years of French instruction and 2–3 years of English instruction at the time of data collection. Pupils in grade 12 (17-18 years old) had had 7 years of French

instruction and 4–5 years of English instruction. In addition to the difference in length of instruction in years, the number of weekly hours also tends to differ: French is generally taught 3–4 hours per week, while English classes take up 2–3 hours of the weekly program.

For more detailed information about the participant characteristics, we refer to Chapter 3 (Section 3.2.).

4.4.3. Instruments

Productive vocabulary tests

Productive vocabulary knowledge was measured using the productive VocabLab tests (Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019). The productive VocabLab tests are form recall tests which require participants to fill in a missing word in a sentence. The vocabulary tests consisted of 60 target items from the 3,000 most frequent words in English and French (more information on the target items can be found in the OSF: https://osf.io/ymkqu/). Each target item was introduced by its first letter (or two letters in case of multiple possible answers), followed by underscores indicating the number of missing letters. At the end of every sentence, we provided a Dutch translation of the target word to minimize alternative responses. The tests had high internal consistency (Cronbach's alpha: English $\alpha = 0.93$, French $\alpha = 0.89$). More detail about the productive VocabLab tests can be found in Chapter 3 (Section 3.3.).

Writing tasks

Participants completed four informal writing tasks, two in English and two in French. We administered two prompts in each language because previous research has shown that the topic can affect learners' output (e.g., Schoonen, 2012; Yoon, 2017; Zenker & Kyle, 2021). The writing prompts were adopted from Schoonen (2012) (see also Schoonen et al., 2003), who developed the prompts in parallel versions to test two languages. Since the prompts were not identical, we could reduce test effects caused by test experience. For the first topic, participants were asked what they would do if they won the lottery (English) or if they could miss school for a year (French). In the other task, they had to complain about their favorite artist not appearing in a magazine (English) or their favorite TV show being canceled (French). More detailed information on the prompt selection and an example of one of the prompts can be found in the OSF: https://osf.io/ymkqu/. For further detail on the writing tasks, see Chapter 3 (Section 3.3.).

4.4.4. Scoring

The vocabulary tests were scored dichotomously (a score of 0 or 1 per item). Responses had to be spelled entirely correctly to be awarded a point. Although this strict scoring system provided a more conservative estimate of learners' vocabulary knowledge, it reflects a level of mastery at which learners can retrieve the correct orthographic form from memory. This was deemed an important aspect of productive vocabulary knowledge. In addition, Johnson et al. (2016) found that test scores based on a strict scoring system were more strongly related to holistic writing scores than lenient scores. This finding suggests it may be useful to measure productive vocabulary knowledge as full *accurate* knowledge of word meaning and form. Both American-English and British-English spelling variants were approved (e.g., *emphasize* and *emphasise*, *honor* and *honour*).

The writing tasks were analyzed in terms of vocabulary use, that is, lexical sophistication and diversity of single words. The two writing tasks in each language were combined and analyzed as a single text. The tasks were anonymized and then analyzed *as is*, meaning no spelling errors were corrected. This was in line with the strict scoring system for the vocabulary tests, wherein incorrectly spelled words were also considered 'not known'. Although incorrectly spelled words may reflect partial knowledge, they also suggest that learners may not have 'full knowledge of meaning and form' (Webb, 2008, p. 85). Since the orthographic form is an important aspect of vocabulary knowledge, it can be argued that these words should not be included in an analysis of learners' L2 vocabulary use.

Lexical sophistication was analyzed by means of a mean-based and a band-based measure: the average log frequency and the lexical frequency profile. We calculated these measures using two subtitle corpora: Subtlex-US for English (Brysbaert & New, 2009) and Lexique (version 3.83) for French (New et al., 2007). Both corpora consist of subtitles from various TV series and films, accounting for a total of 51 and 52 million tokens respectively. For consistency with the lemma-based VocabLab tests, we adopted lemma frequencies for our calculations (e.g., occurrences of *living* and *lived* were counted under the lemma frequency for the verb *live*). For Lexique, the lemma frequencies per million words are publicly available, whereas for Subtlex-US, lemma frequencies have not been reported (only word form frequencies). Therefore, we calculated the lemma frequencies per million words for the Subtlex-US corpus[26]. We then assigned the lemmas in each corpus (ordered by descending frequency) to 1,000-word bands. Lexical sophistication in the learner texts was analyzed by means of (1) average log frequency, by averaging the lemma frequencies of the words

[26] The resulting lemma frequency list is available upon request.

in each learner text (mean-based measure), and (2) lexical frequency profile, by computing the proportion of vocabulary beyond the 2,000 most frequent lemmas ('beyond 2K', excluding off-list words), represented by a single percentage value (band-based measure).

Lexical diversity was measured as the moving-average type-token ratio (MATTR) using a Python package by Kyle et al. (2021). MATTR is calculated by averaging the type-token ratios of equally long contiguous text segments. Consequently, it is particularly stable across text lengths (Kyle et al., 2021; Zenker & Kyle, 2021; Bestgen, 2024), which was important given the text length differences of the participants' writing output. The lemmatized learner texts used for lexical sophistication were used to calculate MATTR for the English and French writing tasks with a window size of 50 words per segment (currently a common window size for written texts).

4.4.5. Data analysis

To compare grade 10 to grade 12 learners (RQ1), we used R (R Core Team, 2023) to run Mann-Whitney U-tests for the form recall scores (total percentage score), lexical diversity values (MATTR), mean-based lexical sophistication (average log frequency), and band-based lexical sophistication (proportion of tokens beyond 2K), for both English and French. A Bonferroni adjustment was applied for multiple testing (α = 0.006). As not all data were normally distributed according to the Shapiro-Wilk test, Mann-Whitney U-tests were computed for all comparisons to facilitate interpretation of the findings.

To answer RQ2 on the predictive power of productive vocabulary knowledge and grade on measures of vocabulary use, we ran a linear mixed-effects model for each dependent variable using the lme4 package in R (Bates et al., 2015). Form recall scores (total percentage score) and grade (grade 10 or 12) were included as predictor variables, and a random effect was added for class group to account for individual variation between the groups. The dependent variables were lexical diversity (MATTR), mean-based lexical sophistication (average log frequency; ALF), and band-based lexical sophistication (proportion of tokens beyond 2K; LFP). For instance, lexical diversity in English was modeled as: lmer(WrEN_S_MATTR ~ VocEN_S*Grade + (1|ClassGroup), data = df). Participants were removed from the model if the standardized residual exceeded 3 SD, which allowed us to exclude outliers from the analyses. We applied a Bonferroni adjustment (α = 0.008). For English, the assumption of linearity was met as shown by the scatter plots and correlations (see Supplementary Data in the published article). Multicollinearity was resolved by centering the data for the vocabulary test, which lowered VIF values well below the threshold

of 5. The Shapiro-Wilk tests did not reveal issues with normality of the residuals, but we were able to improve the Q-Q plots by applying box-cox transformations of the dependent variables. Finally, the residuals vs. predicted plot indicated no heteroscedasticity for any of the variables, indicating that the data could reliably be modeled in a linear mixed-effects model. For French, the assumption of linearity was met only for lexical diversity (see Supplementary Data in the published article), although there were small significant correlations between vocabulary knowledge and mean-based (r_s = .13) and band-based (r_s = .10) lexical sophistication, which were interpreted as a sufficient degree of linearity to justify running the models. We centered the vocabulary test scores to resolve multicollinearity. Shapiro-Wilk tests and Q-Q plots indicated non-normality of the residuals, which was resolved by means of data transformation (square-root transformation for MATTR, log transformation for ALF, box-cox transformation for LFP). Finally, the residuals vs. predicted plot indicated heteroscedasticity for the LFP. For this model, we will thus report robust standard errors to account for this limitation in the model (Cribari-Neto & Lima, 2014).

4.5. Results

4.5.1. Cross-sectional differences in vocabulary knowledge and use

Descriptive results of participants' performance on the vocabulary tests and writing tasks are displayed in Table 4.1. A cross-sectional comparison of the mean form recall scores shows that grade 12 learners outperformed their younger counterparts on the vocabulary test in both languages. In terms of writing output, lexical diversity values (MATTR) appeared to increase slightly from grade 10 to grade 12. Values for mean-based lexical sophistication (ALF) showed a modest decrease between grades for English, indicating lower average frequency and thus more advanced vocabulary use in grade 12, whereas there appeared to be a slight increase for French. Band-based lexical sophistication values (LFP >2K) seemed comparable across grades in both English and French.

The Bonferroni-adjusted results for the Mann-Whitney U-tests for L2 English and French can be found in Table 4.2. (α = 0.006). As suggested by the descriptive findings, grade 12 learners were found to have better vocabulary knowledge at the form recall level than grade 10 learners in both languages (English r = -0.436; French r = -0.544; p <.001). They also included more varied vocabulary in their writing output than their younger counterparts (English r = -0.320; French r = -0.354; p <.001). Specifically for English, mean-based lexical sophistication was found to improve from grade 10 to 12 (r = 0.201; p <.001), as shown by the decreasing mean in Table

4.1. However, no difference in average frequency was found between the grades for French (p = 0.032), indicating that this aspect of vocabulary use in French did not change from one grade to another. Finally, no significant differences were observed between the grades in terms of band-based sophistication in either language (English p = 0.026; French p = 0.023).

Table 4.1. Descriptive results of vocabulary tests and writing output

		English		French	
		Grade 10	**Grade 12**	**Grade 10**	**Grade 12**
Vocabulary knowledge *M (SD)* (%)	1–2K	65.75 (16.37)	78.26 (13.63)	25.48 (14.70)	38.64 (14.80)
	3K	44.71 (19.46)	58.12 (19.97)	13.46 (10.92)	21.28 (11.93)
	Total	55.23 (17.28)	68.19 (16.08)	19.47 (11.96)	29.96 (12.27)
Lexical diversity *M (SD)* (max = 1)	MATTR	0.74 (0.03)	0.76 (0.03)	0.68 (0.05)	0.70 (0.03)
Lexical sophistication *M (SD)*	ALF	3.35 (0.09)	3.32 (0.09)	2.60 (0.12)	2.63 (0.13)
	LFP >2K	0.06 (0.02)	0.07 (0.02)	0.05 (0.02)	0.05 (0.02)

Note. MATTR = moving-average type-token ratio. ALF = average log frequency. LFP >2K = lexical frequency profile, proportion of tokens beyond 2K.
Note. Vocabulary scores N = 441; writing measures N = 423.

Table 4.2. Mann-Whitney U-test results comparing grade 10 and grade 12 participants in terms of vocabulary knowledge, lexical diversity and lexical sophistication

	English			French		
	Statistic	***p***	**Effect Size**	**Statistic**	***p***	**Effect Size**
Vocabulary knowledge	12182	<.001*	-0.436	9852	<.001*	-0.544
MATTR	14669	<.001*	-0.320	13937	<.001*	-0.354
ALF	25929	<.001*	0.201	18933	0.032	-0.123
LFP (>2K)	18843	0.026	-0.127	18770	0.023	-0.130

* p<.006 (significance after Bonferroni adjustment)
Note. Effect size is given by the rank biserial correlation.

As grade 12 learners were not observed to use a larger proportion of words beyond 2K, we also compared the proportions of vocabulary at the 1K and 2K level, as well as off-list words, which include incorrectly spelled and non-existent words like *thereselfs* for *themselves*, or *sûrtenement* for *certainement* (Table 4.3.; α = 0.008). These Mann-Whitney U-tests indicated that for English, there seemed to be little difference between the grades at the 1K and 2K levels. However, grade 10 learners produced significantly more off-list words (r = 0.215). For French, significant cross-sectional differences were observed at the 1K level (r = -0.336) and in the proportion of off-list words (r = 0.381), with grade 10 producing fewer 1K and more off-list words.

Table 4.3. Mann-Whitney U-test results comparing grade 10 and grade 12 participants in terms of band-based lexical sophistication

	English			French		
	Statistic	***p***	**Effect Size**	**Statistic**	***p***	**Effect Size**
LFP 1K	22205	0.617	0.029	14325	<.001*	-0.336
LFP 2K	18787	0.023	-0.130	20050	0.214	-0.071
LFP Offlist	26234	<.001*	0.215	29818	<.001*	0.381

* p<.008 (significance after Bonferroni adjustment)
Note. Effect size is given by the rank biserial correlation.

4.5.2. L2 and L1 vocabulary use

As the writing tasks were based on informal prompts, this informal context may have led learners to (un)consciously employ simpler words. To verify whether the prompts could elicit sophisticated vocabulary, we collected additional data by L1 speakers of English (N = 30) and French (N = 30), who completed the same writing tasks as our L2 participants[27]. A comparison between the L1 and L2 writers showed clear and large differences in sophistication, with L1 writers using significantly more advanced vocabulary than both grade 10 and grade 12 learners (Table 4.4.). Not only did L1 writers write texts with lower average frequency, they also used significantly more low-frequency lemmas beyond 2K. By way of

[27] We collected these additional data via the platforms Prolific (L1 English) and Clickworker (L1 French). Due to age restrictions of these platforms, the minimum age of the L1 participants was 18 years old. To keep the L1 sample as comparable as possible to the L2 sample, we set the age limit to 20 years old. Prolific and Clickworker members could only participate if they met three criteria: (1) their L1 was English (Prolific) or French (Clickworker), (2) they were between 18 and 20 years old, (3) they were a student in secondary school or an undergraduate program. The mean age of the L1 participants was 19 years old. Participants were explicitly instructed not to leave the online writing environment during the task.

illustration, some examples of words in English used only by grade 10 and grade 12 learners were *scary, natural, village, surprised, recognize,* and in French *lourd, endormir, sec, étude, frais*. Words used only by L1 English writers included *frivolous, exhilarating, audacity, atrocious, exacerbate,* and by L1 French writers *ensommeillé, diversifier, insoupçonné, rebondissement, enrichissant.*

Table 4.4. Mann-Whitney U-test results comparing output by L2 and L1 writers

	English			French		
	Statistic	***p*-value**	**Effect Size**	**Statistic**	***p*-value**	**Effect Size**
Grade 10 – L1						
MATTR	614	<.001*	-0.762	341	<.001*	-0.868
ALF	4668	<.001*	0.809	4939	<.001*	0.914
LFP >2K	675	<.001*	-0.738	786	<.001*	-0.695
Grade 12 – L1						
MATTR	1760	<.001*	-0.533	875	<.001*	-0.768
ALF	6373	<.001*	0.693	7273	<.001*	0.932
LFP >2K	1236	<.001*	-0.672	1279	<.001*	-0.660

* p<.004 (significance after Bonferroni adjustment)
Note. Effect size is given by the rank biserial correlation.

4.5.3. Vocabulary knowledge and grade as predictors of vocabulary use

In order to illustrate the descriptive relationships between the variables, Table 4.5. shows the Spearman correlations between the four vocabulary measures: vocabulary knowledge (form recall), lexical diversity (MATTR), mean-based lexical sophistication (ALF), and band-based lexical sophistication (LFP).

Table 4.5. Spearman correlation matrix of the vocabulary measures

	Form recall	**MATTR**	**ALF**	**LFP (>2K)**
English				
Form recall	1			
MATTR	0.48	1		
ALF	-0.42	-0.57	1	
LFP (>2K)	0.28	0.40	-0.77	1

French				
Form recall	1			
MATTR	0.44	1		
ALF	0.13	0.03	1	
LFP (>2K)	0.1	0.14	-0.05	1

The linear mixed-effects models for L2 English (Table 4.6.) showed that vocabulary knowledge was a significant predictor of lexical diversity, mean-based and band-based lexical sophistication, with small- to medium-sized effects. The models explained 28% of the variance in diversity and 20% of the variance in mean-based sophistication, but only 9% of the variance in band-based sophistication. Grade did not significantly predict any of the measures of vocabulary use, nor did it interact with vocabulary knowledge in predicting vocabulary use (see Supplementary Data in the published article for the interaction plots). The random effect of class group contributed only minimally to the models.

Table 4.6. Mixed-effects models for English with form recall scores and grade as predictor variables

Dependent variable	**Conditional R^2**	**Random effects[a]**	***β (SE)***		
			Form recall	**Grade**	**Form recall*Grade**
MATTR	0.28	0.213e-4	0.419*** (0.009e-2)	0.251 (0.003)	0.009 (0.012e-2)
ALF	0.20	0.002	-0.394*** (0.001)	-0.072 (0.033)	-0.028 (0.002)
LFP (>2K)	0.09[b]	0.000	0.247** (0.019e-2)	0.038 (0.004)	0.067 (0.025e-2)

*** p<.001; ** p<.008 (significance after Bonferroni adjustment)
[a]Random effect for class group (N = 27).
[b]Marginal R^2 is reported because the variance of the random effects was estimated to be zero (singularity).

Further, the linear mixed-effects models for L2 French (Table 4.7.) revealed that both vocabulary knowledge and grade significantly predicted lexical diversity, showing medium-sized effects. Grade also interacted with vocabulary knowledge in predicting lexical diversity, with a stronger impact of vocabulary knowledge on lexical diversity in grade 10 than grade 12 (see Supplementary Data in the published article). The French models explained 30% of variance in diversity, while none of the predictors

significantly explained the variance in sophistication. Grade also did not interact with the relationship between vocabulary knowledge and lexical sophistication as measured by average frequency or LFP. The contribution of the random effect of class group was minimal.

Table 4.7. Mixed-effects models for French with form recall scores and grade as predictor variables

Dependent variable	Conditional R^2	Random effects[a]	*β (SE)*		
			Form recall	Grade	Form recall*Grade
MATTR	0.30	0.405e-4	0.495*** (0.013e-2)	0.400** (0.003)	-0.292** (0.017e-2)
ALF	0.09	0.686e-4	0.112 (0.021e-2)	0.135 (0.005)	0.017 (0.027e-2)
LFP (>2K)	0.05	0.603e-5	0.022 (0.011e-2)	0.223 (0.002)	0.027 (0.014e-2)

*** p<.001; ** p<.008 (significance after Bonferroni adjustment)
[a]Random effect for class group (N = 27).
Note. Robust standard errors for LFP (>2K).

4.6. Discussion

4.6.1. Cross-sectional differences in vocabulary knowledge and use

RQ1 aimed to investigate the cross-sectional differences in vocabulary knowledge and use between grade 10 and grade 12. We compared the results of the form recall test, lexical diversity and lexical sophistication, using both a mean-based and band-based measure of sophistication for more detailed insight into learners' vocabulary use. Vocabulary knowledge at the form recall level improved from grade 10 to 12 in both English and French, corroborating previous research on receptive vocabulary (Peters, Noreillie, et al., 2019; Webb & Chang, 2012) and productive vocabulary (Laufer, 1998; Henriksen & Danelund, 2015), and suggesting that, as could be expected, having additional years of formal instruction was positively related to L2 vocabulary knowledge. Writing output by grade 12 learners was also more lexically diverse than that by grade 10 learners in both languages (Henriksen & Danelund, 2015). Specifically for English, the average frequency of the learners' texts decreased with grade, pointing to more advanced vocabulary use by grade 12 learners (Consolini & Kyle, 2024; Crossley et al., 2013; Crossley &

McNamara, 2012; L. Guo et al., 2013; Kim et al., 2018b). In contrast, mean-based lexical sophistication did not differ cross-sectionally for French, which seems to imply that grade 12 students' additional L2 French instruction did not result in more advanced vocabulary. Likewise, band-based lexical sophistication was not found to improve between grades for either language, in line with other studies that have demonstrated a lack of increase in band-based sophistication across grades in secondary school (Laufer, 1998; Levitzky-Aviad & Laufer, 2013; Henriksen & Danelund, 2015).

One possible explanation is that lexical sophistication does not progress until a more advanced stage of L2 acquisition, which could explain why Levitzky-Aviad and Laufer (2013) only found a significant increase in advanced vocabulary use at the university level. For English, we found a small but significant cross-sectional difference in the average frequency of the learners' texts, which was not reflected in the proportion of lemmas beyond 2K. Due to the nature of the condensed LFP (1,000-word bands with a rather arbitrary cut-off at 2K), this measure may not have been able to capture the small advances in sophisticated vocabulary use that seem to have taken place from grade 10 to 12 (see also Crossley et al., 2013). For French, on the other hand, neither the mean-based nor the band-based measure detected any difference between the grades. Given the low form recall scores for French, even in grade 12, learners' French vocabulary knowledge may have been too limited to allow for more advanced vocabulary use in their writing. This interpretation is supported by the observed increase in 1K words and decrease in off-list words, which may indicate that grade 12 participants produced fewer non-existent words and were able to use more existing (high-frequency) words. Additionally, learners' limited vocabulary knowledge may also explain the small correlation coefficients between French form recall knowledge and measures of sophistication: a linear relationship may not yet be present due to little variation in form recall scores.

Another aspect that could have affected learners' lexical sophistication are the writing tasks. First, prompts have been shown to affect learners' vocabulary use (e.g., Yoon, 2017; Zenker & Kyle, 2021). The informal prompts may have promoted more high-frequency words. However, the comparison with the L1 data showed a large difference in advanced vocabulary use between L1 and L2 writers in both English and French (Table 4.4.). It should be noted that the age and maturity of the L1 participants (M_{age} = 19) might have affected their vocabulary use to some extent. Nevertheless, the fact that L1 writers used more sophisticated vocabulary in response to the same prompts, suggests that our prompts were able to elicit sophisticated vocabulary, and the limited increase in sophistication from grade 10 to 12 cannot solely be attributed to the prompts. Second, the findings for sophistication may also be related to the

task instructions, which did not state that participants were expected to use sophisticated words. As a result, it is possible that our writing tasks did not test *ability* as such, but rather *typical performance* in terms of lexical sophistication (Cronbach, 1970). Learners might have been able to use advanced vocabulary 'had they known' this was part of the scoring (Schoonen, 2011), but this ability was not reflected in their output because they were unaware their vocabulary use would be analyzed. The higher levels of sophistication by L1 writers do not necessarily invalidate this explanation: given the higher proficiency of L1 writers, the use of advanced vocabulary may be more automatized. Therefore, it is possible that L1 writers' output also reflects their typical behavior, but the use of sophisticated words may simply require a less conscious effort for them than for L2 learners.

4.6.2. Vocabulary knowledge and grade as predictors of vocabulary use

RQ2 aimed to examine the extent to which vocabulary knowledge and grade could predict vocabulary use. First, the mixed-effects models for English and French both showed that vocabulary knowledge at the form recall level significantly predicted lexical diversity. This indicates that learners with better productive vocabulary knowledge tended to employ a larger variety of words in their writing, in line with Henriksen and Danelund (2015). This finding may seem self-evident based on research that has shown that more proficient learners produce more lexically diverse output (Engber, 1995; Treffers-Daller et al., 2018; Zenker & Kyle, 2021). Nonetheless, even if learners have more vocabulary knowledge available, they might not necessarily use all of the vocabulary in their mental lexicon and instead choose to repeat the same words multiple times. This was not the case in our study: participants with higher vocabulary scores were more likely to avoid repetition.

With regard to lexical sophistication in English, the mixed-effects models showed that vocabulary knowledge was a significant predictor of mean- and band-based sophistication, confirming some previous studies (Henriksen & Danelund, 2015; Johnson et al., 2016; Karakoç & Köse, 2017). In contrast, our results for French were in line with Laufer (1998), in that no such relationship was established for either sophistication measure. It is worth mentioning that this may be due to the low level of linearity between vocabulary knowledge and lexical sophistication measures (see Data analysis). However, this difference may also suggest an impact of the L2 learning setting on the relationship between vocabulary knowledge and use. Flemish L2 learners are taught French in school at an earlier age than English, yet they are rarely exposed to French out-of-school activities. With

Flemish learners being exposed to large amounts of out-of-school English from a young age (e.g., De Wilde et al., 2020b; Puimège & Peters, 2019a) and having more favorable attitudes towards English than French (e.g., Dewaele 2005b), this creates a distinct learning setting for these two L2s, even if they are both taught as mandatory subjects in school. The learning setting of Laufer's (1998) participants was similar to the one for L2 French learners in Flanders in that it presents a low-exposure context with most L2 input from formal instruction. Given the positive relationship between out-of-school L2 contact and vocabulary knowledge (e.g., De Wilde et al., 2022; Peters, Noreillie, et al., 2019; Puimège & Peters, 2019a), this may help to explain the lack of a relationship for French.

Alternatively, another explanation for this difference may also be found in the limited French vocabulary knowledge of Flemish learners. Considering the low form recall scores for French, participants' vocabulary knowledge may have been too limited to detect any predictive power of vocabulary knowledge on lexical sophistication. Nevertheless, our lemma-based analyses might have also affected these findings, since French is a more morphologically rich language, and lemmas in French contain more distinct inflected forms than in English (e.g., French verb *vivre* in the present and past simple: *vis* and *vécu*, contrary to the simpler English forms *live* and *lived*). While it is clear that our participants had a low level of vocabulary knowledge, the lemma-based measures of the writing output might underestimate learners' French vocabulary use to some degree.

In addition to the role of vocabulary knowledge in vocabulary use, we also looked at its interaction with grade. Grade only interacted with vocabulary knowledge in predicting lexical diversity in French, with a larger effect of vocabulary knowledge on lexical diversity in grade 10. This may reveal the critical role of vocabulary knowledge at a less advanced stage in L2 acquisition. An increase in vocabulary knowledge, as demonstrated by the medium effect size in the cross-sectional comparison of grade 10 and 12, would then result in a less determining role of vocabulary knowledge in lexical diversity. Grade was not found to predict any of the other measures, nor did it interact with vocabulary knowledge in predicting other measures of vocabulary use in either English or French.

4.6.3. Pedagogical implications

The findings of the present study have pedagogical implications. Based on the cross-sectional results, it seems that the L2 learners did not strongly improve their use of advanced vocabulary in English or French. Although the average frequency did seem to somewhat increase from grade 10 to 12 in English, no difference was observed for average frequency in French, or for the condensed LFP in either language. Since there was a

moderate increase in form recall knowledge and lexical diversity as opposed to lexical sophistication, this may point to a need for more attention to advanced vocabulary in the L2 classroom. As Levitzky-Aviad and Laufer (2013) suggest, learners in secondary school may lack sufficient practice in using advanced vocabulary, which could lead learners to adopt a risk-avoiding strategy (Laufer, 1998). Learners should instead be prompted to use advanced vocabulary, so they do not rely solely on words familiar to them.

In addition, it seems realistic that L2 writing tasks place a substantial cognitive load on learners, especially if they have to contemplate both diversity and sophistication of their vocabulary. In that case, learners may prioritize certain aspects of vocabulary use, which could explain why sophistication was not always observed to increase when diversity was. L2 learners, particularly those with less vocabulary knowledge, may benefit from a more phased approach to improving their vocabulary use in writing. If learners are guided through the process of learning to use (varied and) advanced vocabulary in smaller steps, the cognitive load may become more manageable, which may in turn lead to higher-quality vocabulary use in their writing output.

Furthermore, the findings indicated that the relationship between vocabulary knowledge and use can depend on the L2 measured. For English, vocabulary knowledge was a predictor of both mean-based and band-based lexical sophistication, whereas vocabulary knowledge did not predict either measure of lexical sophistication in French. This difference may imply that pedagogical strategies should be tailored to the learner group, taking into account their L2 proficiency and the learning setting (e.g., out-of-school L2 contact). For instance, learners with better vocabulary knowledge who are often exposed to the L2 may benefit from instruction challenging them to practice low-frequency words. By contrast, for learners with limited vocabulary knowledge and out-of-school exposure, it may be necessary to adopt a different focus in the L2 classroom to guide learners towards increased lexical proficiency in smaller steps.

4.6.4. Limitations and future directions

There are a few limitations to this study that should be mentioned. First, vocabulary use was operationalized as lexical diversity and lexical sophistication. The study did not take into account other measures of lexical proficiency, like accuracy (use of the correct form) and appropriateness (use of semantically correct vocabulary suitable to the topic and register). Likewise, the learner texts were not holistically scored by raters. In future research, analyses of accuracy and appropriateness as well as holistic ratings would serve as a valuable addition to data on lexical

diversity and sophistication. Related to accuracy, we adopted a strict scoring system wherein the writing output was not corrected prior to the analyses. A lenient coding system which rewards learners for partial knowledge could provide a different perspective on L2 vocabulary use in writing. This study also focused on an analysis of single words only. Given that the use of multi-word units has emerged as an important predictor of L2 writing performance (e.g., Bestgen & Granger, 2014; Kim et al., 2018; Siyanova-Chanturia, 2015), the use of formulaic language in L2 English and French may be an interesting topic of follow-up studies.

Second, this study examined vocabulary use in two L2s with a view to analyzing patterns across languages. However, a direct cross-linguistic comparison may not be warranted due to the morphological differences between English and French. In future research, authors might consider comparing orthographic types rather than lemmas, or they could apply an alternative lemmatization scheme to account for the morphological richness of French (for an example with Spanish, see Consolini & Kyle, 2024). Similarly, different findings might emerge for future studies that operationalize lexical indices in alternative ways, for instance, by manipulating the window size, employing other measures of vocabulary use, or using different reference corpora (see also Bottini & Le Foll, 2024).

4.7. Conclusion

The present study investigated the lexical proficiency of L2 learners in two languages (English and French) and two grades (grade 10 and 12). Participants completed a form recall vocabulary test and two writing tasks in English and French. Vocabulary use was operationalized as lexical diversity and both mean-based and band-based sophistication. The cross-sectional comparison showed that vocabulary knowledge and lexical diversity improved from grade 10 to 12, whereas mean-based lexical sophistication only increased across grades in English. Band-based sophistication was not found to increase between grades in either language. Further, the analyses showed that productive vocabulary knowledge was a predictor of lexical diversity in writing for both English and French, while vocabulary knowledge could predict both measures of lexical sophistication in English but not in French.

This study illustrated the relatively limited use of advanced vocabulary in L2 learners' written output across languages and across grades. While lexical diversity improved for both languages with more formal instruction and productive vocabulary knowledge, our findings for lexical sophistication point to differences between the L2s, possibly depending on learners' proficiency level and the learning setting. Since

learning how to use advanced vocabulary in writing remains an important aspect of L2 vocabulary and writing development, we conclude that it may be necessary to attach greater importance to the development of lexical sophistication within the L2 classroom.

5. The role of out-of-school contact and instruction in productive vocabulary knowledge and vocabulary use[28]

Abstract

This study examined the productive vocabulary knowledge and written vocabulary use of second language (L2) learners of English and French, in relation to their out-of-school L2 contact and length of formal instruction. Data were collected from L2 learners who completed a form recall vocabulary test, two writing tasks and a questionnaire in both L2 English and L2 French. The results showed that certain types of out-of-school L2 contact with English (watching TV with L2 captions, speaking and writing, reading, YouTube, gaming) were associated with learners' productive vocabulary knowledge and vocabulary use in writing, measured as lexical diversity (moving-average type-token ratio) and lexical sophistication (average log frequency and lexical frequency profile). In addition, length of instruction was also an important predictor of learners' productive vocabulary knowledge and use, especially for L2 French where out-of-school contact was limited. The findings demonstrate the role of input in L2 learning, indicating that out-of-school L2 contact and formal instruction can serve as valuable sources of L2 contact in the development of learners' productive vocabulary knowledge, as well as their lexical diversity and sophistication in writing.

Keywords: out-of-school contact; length of instruction; L2 English; L2 French; productive vocabulary knowledge; lexical diversity; lexical sophistication

[28] Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., & Peters, E. (under review). The role of out-of-school language contact and instruction in L2 English and L2 French vocabulary knowledge and use.

5.1. Introduction

In recent years, research has increasingly investigated the role of out-of-school language contact in second language (L2) learning. Out-of-school language contact, through activities like watching TV, gaming and reading, can serve as an important source of L2 input for learners. Usage-based theories in particular have emphasized the importance of L2 input for L2 learning (e.g., Ellis & Collins, 2009; Tomasello, 2003). Indeed, previous studies have consistently demonstrated a positive relationship between out-of-school L2 contact and receptive L2 vocabulary knowledge (e.g., González-Fernández & Schmitt, 2015; Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019), even before the start of formal L2 instruction (e.g., De Wilde et al., 2020b; Puimège & Peters, 2019a). Yet, with a few exceptions, less is known about out-of-school L2 contact in relation to productive vocabulary knowledge (e.g., Bollansée et al., 2021; Sundqvist, 2019). Even fewer studies have investigated the role of out-of-school contact in lexical diversity (e.g., Kaatari et al., 2023) and lexical sophistication (e.g., Sundqvist & Wikström, 2015) in L2 writing. Additionally, findings on the role of specific out-of-school L2 activities have not always been consistent.

The majority of research has investigated out-of-school contact in the context of L2 English (Kusyk et al., 2025). Two exceptions are De Wilde et al. (2022) and Peters, Noreillie, et al. (2019), whose comparisons of L2 English and L2 French suggested that out-of-school contact may only benefit L2 proficiency if learners are exposed to large amounts of L2 input. However, their studies focused on receptive vocabulary knowledge only. No studies to our knowledge have investigated the role of out-of-school contact in productive vocabulary knowledge and use for multiple L2s. Given that English is the language of pop culture (Werner & Tegge, 2021) and possibilities for out-of-school contact may be less extensive for other L2s, more research is warranted on the role of out-of-school contact in L2 proficiency in other languages than English.

The present study aims to investigate the relationship between out-of-school L2 contact and two underexplored areas of L2 proficiency, namely, productive vocabulary knowledge and written vocabulary use. Vocabulary use was operationalized as lexical diversity (use of varied vocabulary) and lexical sophistication (use of low-frequency vocabulary). In addition, given that formal instruction is another important source of input for L2 learners (e.g., Peters, Noreillie, et al., 2019), the study aims to shed light on the role of length of instruction in the development of productive vocabulary knowledge, lexical diversity and sophistication. The study focuses on two L2s, English and French, in order to advance our understanding of how out-of-school contact and formal instruction may affect productive vocabulary knowledge and lexical diversity and sophistication in two L2s.

5.2. Literature review

5.2.1. Out-of-school L2 contact and vocabulary knowledge

Research has shown the beneficial role of out-of-school language contact[29] in L2 vocabulary learning. Learners who are often exposed to L2 English outside of the classroom tend to know more words receptively, which holds true for learners prior to formal L2 instruction (De Wilde et al., 2020b; Puimège & Peters, 2019a) and after the onset of instruction (Hannibal Jensen, 2017; Sundqvist, 2019). Some studies have suggested that out-of-school contact may be a stronger predictor of L2 vocabulary than formal instruction (Busby, 2021; Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024; Warnby, 2022). This is in keeping with usage-based theories of language learning, which posit that languages are learned through repeated exposure and use, emphasizing the importance of extensive L2 input (e.g., Ellis & Collins, 2009; Tomasello, 2003).

Previous research has also looked at how learners' engagement with out-of-school L2 activities may affect their vocabulary knowledge. A common activity among L2 learners is watching TV in the L2. Research has found TV viewing to be a strong predictor of receptive L2 vocabulary knowledge (González-Fernández & Schmitt, 2015; Hannibal Jensen, 2017; Kuppens, 2010; Peters & Webb, 2018; Rodgers, 2013). Further, L2 contact often takes place online, through activities like gaming, watching YouTube videos and using social media. Especially gaming has repeatedly been established as a predictor of receptive L2 vocabulary knowledge (Hannibal Jensen, 2017; Kuppens, 2010; Peters, Noreillie, et al., 2019; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Research has also shown that boys tend to game more frequently than girls, which is often accompanied by better L2 vocabulary knowledge (Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Indeed, Peters, Noreillie, et al. (2019) found through structural equation modeling that gender affected out-of-school L2 contact, which in turn affected receptive L2 vocabulary knowledge.

Two more activities deserve to be discussed. First, listening to music is a popular activity among L2 learners, although it has not always emerged as a positive predictor of L2 vocabulary knowledge (e.g., Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019). This is likely related to the nature of listening to L2 music: although learners may spend much time listening to music, they may not be actively engaged in the activity and therefore not

[29] Out-of-school L2 contact is defined as any contact with the L2 initiated by the learner, outside of and unrelated to the L2 classroom. These learner-initiated L2 activities are also commonly referred to as 'extramural exposure', or specifically for English 'extramural English' (Sundqvist, 2024).

benefit from the additional L2 input. Nevertheless, some studies have demonstrated a positive correlation between music and receptive L2 vocabulary knowledge (e.g., Puimège & Peters, 2019a). In contrast with music, reading seems to be one of the least popular activities among L2 learners. The low levels of L2 contact through reading are likely an important reason that most studies have not found a relationship between out-of-school reading and vocabulary knowledge (Arndt, 2023; Bollansée et al., 2021; Hannibal Jensen, 2017; Kuppens, 2010; Peters, Noreillie, et al., 2019; Puimège & Peters, 2019a). This may also be related to the way in which out-of-school contact is measured (e.g., Likert scales), which may lead to limited variation in the data for this activity.

Although many studies have examined the relationship between out-of-school L2 contact and receptive vocabulary knowledge, only limited research is available on the role of out-of-school contact in productive vocabulary knowledge, and the findings are mixed. Some studies have shown a positive relationship between out-of-school L2 contact and productive vocabulary knowledge (Bollansée et al., 2021; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Other studies, however, have indicated that the benefits of out-of-school contact may not always extend to a productive level of mastery (Kuppens, 2010; Schwarz & Ghamarian-Krenn, 2024).

Research on the relationship between specific out-of-school activities and productive vocabulary knowledge is also inconclusive. For instance, Kuppens (2010) found a positive relationship between TV viewing with L1 subtitles and L1-to-L2 translation scores, while Bollansée et al. (2021) established a negative correlation between TV with L1 subtitles and form recall scores. Bollansée et al. (2021) did, however, find a positive effect of non-subtitled TV, supporting previous studies on receptive vocabulary knowledge (De Wilde et al., 2020b; Peters & Webb, 2018). Additionally, the majority of studies on out-of-school contact and productive vocabulary knowledge have focused on the beneficial effect of gaming. With one exception (Kuppens, 2010), these studies have established a positive correlation between gaming and productive vocabulary knowledge in English, both before and after the onset of instruction (Bollansée et al., 2021; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012).

Finally, it should be noted that previous studies on out-of-school L2 contact have focused almost exclusively on L2 English. To our knowledge, only two exceptions have investigated the impact of out-of-school contact on vocabulary in multiple L2s (De Wilde et al., 2022; Peters, Noreillie, et al., 2019). Both studies demonstrated a positive relationship between out-of-school English and receptive vocabulary knowledge of primary school (De

Wilde et al., 2022) and secondary school and university learners (Peters, Noreillie, et al., 2019), whereas there was only a limited impact of out-of-school French on French vocabulary knowledge. Given the potential differences in out-of-school contact with L2 English and L2 French, it remains to be seen to what extent out-of-school contact may affect productive vocabulary knowledge differently in L2 English and French.

5.2.2. Out-of-school L2 contact and vocabulary use in writing

An underrepresented aspect of L2 proficiency in research on out-of-school contact, is vocabulary use in L2 writing. Vocabulary use can be analyzed in terms of lexical diversity (variety of words in a text) and lexical sophistication (use of low-frequency words).

Only two studies, as far as we know, have focused on lexical diversity, both with Swedish learners of L2 English. A small-scale analysis ($N = 16$) by Sundqvist (2019) seemed to indicate a positive effect of out-of-school contact on learners' lexical diversity. However, Sundqvist's study used the type-token ratio as a measure of lexical diversity, which has been criticized for its dependence on text length. More recently, Kaatari et al. (2023) analyzed lexical diversity as the moving-average type-token ratio (MATTR), which is stable across text lengths (e.g., Zenker & Kyle, 2021). The participants completed an argumentative essay and a questionnaire gauging their weekly contact with L2 English. Kaatari et al.'s study confirmed that L2 learners who had conversations in English and watched English-language TV more often, produced more lexically diverse texts. However, in contrast with findings on productive vocabulary knowledge, gaming was not related to lexical diversity.

Further, only few studies have examined the impact of out-of-school contact on lexical sophistication. Sundqvist and Wikström (2015) and Sundqvist (2019) found that gaming in L2 English was related to the use of more polysyllabic words in writing. Nevertheless, polysyllabic words are not necessarily low-frequent (e.g., *family, experience, easier, listening*). Two other studies have analyzed lexical sophistication in L2 English based on word frequency, and established no correlations between out-of-school L2 contact and low-frequency or academic vocabulary use in L2 writing (Cobb & Horst, 2011; Olsson & Sylvén, 2015). This may in part be due to the measure of lexical sophistication used, that is, the proportion of low-frequency vocabulary in learner texts based on 1,000-word bands. Such a band-based measure may not always capture small differences in low-frequency vocabulary use (Caltabellotta et al., 2025). Given the limited research on the role of out-of-school L2 contact in learners' diversity and sophistication in L2 writing, further research is warranted.

Finally, research on the relationship between out-of-school contact and lexical diversity and sophistication has exclusively focused on English. No studies to our knowledge have investigated the role of out-of-school L2 contact in lexical diversity and sophistication in multiple L2s, limiting the generalizability of findings.

5.2.3. Formal language instruction

Out-of-school contact has become a common form of input outside of the L2 classroom, and for L2 English it often precedes formal instruction (Sundqvist, 2024). Yet, not all learners have large amounts of out-of-school L2 contact, requiring them to rely mostly on formal classroom instruction. Formal instruction is a significant predictor of receptive vocabulary knowledge, as demonstrated with primary school (e.g., Unsworth et al., 2014), secondary school and university students (e.g., Laufer, 1998; Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024; Webb & Chang, 2012). In a longitudinal study, Webb and Chang (2012) found that participants' receptive vocabulary knowledge increased yearly throughout secondary and tertiary education, although the gains were modest. Length of instruction was also a significant predictor of receptive vocabulary knowledge in several cross-sectional studies (Laufer, 1998; Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024), while other studies have found only weak or non-significant correlations between instruction and receptive vocabulary (Busby, 2021; De Wilde et al., 2022). Peters, Noreillie, et al. (2019) demonstrated that length of instruction predicted receptive vocabulary knowledge in two L2s, English and French, although the impact was limited if instruction was not complemented with out-of-school contact.

Only a few studies have examined the impact of length of instruction on learners' productive vocabulary knowledge. In a cross-sectional design, Laufer (1998) demonstrated that L2 learners of English who had had additional years of instruction showed better productive vocabulary knowledge. Caltabellotta et al. (2025) corroborated this finding for both L2 English and French. In addition, Schwarz and Ghamarian-Krenn (2024) found that length of instruction was the only predictor of scores on a word association task, suggesting the importance of formal instruction for productive vocabulary knowledge.

Research on the role of formal instruction in written vocabulary use is also limited, with a few exceptions. Caltabellotta et al. (2025) found that grade 12 learners produced more lexically diverse texts than grade 10 learners in both L2 English and French. Laufer (1998) analyzed lexical sophistication in L2 English texts by grade 10 and grade 11 learners, and detected no difference in their low-frequency word use through a band-

based approach (lexical frequency profile). Caltabellotta et al. (2025), however, measured low-frequency vocabulary use through both a band-based (lexical frequency profile) and count-based (average log frequency) measure. They found that the count-based measure picked up on a small increase in low-frequency word use in English between grade 10 and grade 12 learners, whereas no difference was observed with the band-based measure. Given the limited research into the relationship between formal instruction and lexical diversity and sophistication, more research is needed using multiple indices of vocabulary use.

5.3. Rationale and research questions

Research on out-of-school contact has repeatedly highlighted the potential of out-of-school L2 contact for L2 learning, mainly focusing on receptive vocabulary knowledge. Yet, questions remain about the benefits of out-of-school L2 contact for productive vocabulary knowledge and lexical diversity and sophistication in L2 writing, aspects of L2 proficiency which have received considerably less attention in previous research. Since learners are increasingly engaging in out-of-school L2 activities, especially in English, it seems relevant to investigate their potential for productive vocabulary knowledge and vocabulary use. Further, more research is needed to determine which out-of-school activities may be most beneficial to the development of L2 learners' productive vocabulary knowledge, lexical diversity and sophistication. In addition, given the omnipresence of L2 English in many European countries, results may differ for other L2s. Yet, with the exception of two studies (De Wilde et al., 2022; Peters, Noreillie, et al., 2019), research on out-of-school contact has rarely examined other languages than L2 English.

The present study aims to address these gaps by shedding more light on the relationship between out-of-school L2 contact and productive L2 vocabulary knowledge on the one hand, and lexical diversity and sophistication in L2 writing on the other. Only one study on this topic has measured lexical diversity using the moving-average type-token ratio (MATTR; Kaatari et al., 2023). Two studies have operationalized sophistication in terms of low-frequency or academic vocabulary use (Cobb & Horst, 2011; Olsson & Sylvén, 2015), both using a band-based approach, while no studies to date have used a count-based approach. In addition to out-of-school contact, this study also addresses length of instruction, as previous studies have demonstrated positive effects on L2 vocabulary knowledge (e.g., Laufer, 1998; Peters, Noreillie, et al., 2019; Webb & Chang, 2012). Finally, we take gender into account as a control variable, as some studies have found gender differences in out-of-school L2 activities (e.g., Sundqvist, 2019). Finally, given the limited research on

languages other than English (De Wilde et al., 2022; Peters, Noreillie, et al., 2019), the focus is on two L2s: English and French.

In sum, the present study examines the relationships between out-of-school L2 contact, length of instruction, gender, and productive vocabulary knowledge and lexical diversity and sophistication in written L2 English and L2 French. The following research questions guided the study:

RQ1. What is the relationship between amount of time spent on out-of-school contact, length of formal instruction, gender, and productive vocabulary knowledge and lexical diversity and sophistication in written L2 English and L2 French?

RQ2. Which type of out-of-school activities are most strongly related to productive vocabulary knowledge and lexical diversity and sophistication in written L2 English and L2 French?

5.4. Methodology

5.4.1. Research context and participants

The present study was conducted in Flanders, the Dutch-speaking region of Belgium. We originally collected data from 441 participants. However, we only included data from participants who completed all tasks in both languages (see below). Further, participants who reported having English or French as a home language were removed from the dataset, so these home languages could not be confounded with out-of-school contact. Finally, we excluded participants from the sample if they had produced fewer than 50 tokens on the writing tasks in at least one language. As a result, the final sample (N = 388) consisted of 152 grade 10 learners (M_{age} = 15.15) and 236 grade 12 learners (M_{age} = 17.09). The majority of the participants identified as female (n = 230), while most others identified as male (n = 153). Five participants reported being non-binary or did not disclose their gender. For more information about the participants, we refer to Chapter 3 (Section 3.2.).

5.4.2. Design and procedure

Data were collected by the first author during school hours. We adopted a within-subject design in which each participant completed a form recall vocabulary test, two writing tasks and a questionnaire in both L2 English and L2 French (see Instruments). We organized two separate sessions (one for each L2) to avoid fatigue effects on the learners'

performance. The order of the sessions was counterbalanced to reduce order effects. Participants completed the tasks via the online platform Gorilla Experiment Builder (www.gorilla.sc), without any resources. Participants were allowed to spend a maximum of 30 minutes on each vocabulary and writing task. No time limit was set for the questionnaires, but participants were typically able to complete each questionnaire in approximately 10 minutes.

5.4.3. Instruments

Learner questionnaires

To collect information on learners' out-of-school contact and length of instruction, we administered a questionnaire for each L2, adapted from previous studies in Flanders (Prophète et al., 2024; Puimège & Peters, 2019a). The questionnaires were provided in Dutch, and the researcher was present during administration to answer any questions. Participants were encouraged to respond as honestly and accurately as possible. The questionnaires are available in the OSF: https://osf.io/bdmwh/.

To gauge learners' out-of-school contact with L2 English and French, participants indicated how much time they spent on average on 11 different L2 activities, by providing an estimate in hours and minutes on a weekly basis. We opted for an average estimate per week so participants would not use a specific day as a reference, which could have led to biased responses. Following previous research on out-of-school contact (e.g., De Wilde et al., 2021; Peters, Noreillie, et al., 2019; Sundqvist, 2019), we included the following L2 activities: watching TV (without subtitles, with L1 subtitles, with L2 captions, with subtitles in another language); watching YouTube videos; gaming; listening to music; reading; listening to or reading social media posts; and speaking or writing in the L2 (to communicate with friends or family, or to create their own content). The questionnaire also included open-ended items where learners could elaborate on their L2 contact (e.g., the kind of games they played), which were beyond the scope of the present study.

With regard to length of instruction, we asked participants in which grade they had received their first English or French class in school. This response was used to calculate the number of years of formal instruction. For instance, if a grade 10 student reporting having had their first English class in grade 7, this was interpreted as 3 years of formal English instruction. Other questions about formal instruction were not analyzed for the present study as they were not relevant to our research questions. A detailed description of the questionnaires can be found in Chapter 3 (Section 3.3.).

Productive vocabulary tests

To test productive vocabulary knowledge, we administered the productive VocabLab tests (Caltabellotta et al., 2025; Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019). The VocabLab tests were developed in parallel versions for English and French, based on the same sampling criteria and following the same format. They are frequency-based form recall tests requiring participants to complete a missing word in a sentence. The first letter(s) of the target word is always indicated, as well as the number of missing letters and an L1 (Dutch) translation of the target word. The VocabLab tests contain target items from the 3,000 most frequent words in English and French, with 30 items from the 1K-2K frequency bands and 30 items from the 3K frequency band. Both tests had high reliability (Cronbach's alpha for English: $\alpha = 0.93$, for French: $\alpha = 0.87$). Detailed information about the productive VocabLab tests is provided in Chapter 3 (Section 3.3.).

Writing tasks

To analyze learners' vocabulary use in writing, we asked the participants to complete two informal writing tasks in each language. Two writing tasks were administered because research has shown that a writing topic can in itself affect vocabulary use (e.g., Zenker & Kyle, 2021). For each language, learners completed one writing task in which they described what they would do if they won the lottery (English) and if they were allowed to miss school for a year (French), and one writing task in which they wrote a complaint about their favorite artist not appearing in a magazine (English) and their favorite TV show being canceled (French). The topics were adapted from Schoonen (2012) (see also Schoonen et al., 2003), who developed the prompts in two parallel versions, eliciting similar output across the two languages. Given the focus on informal writing, the recipient of the writing tasks was a friend. Instructions were given in the L1 (Dutch) to avoid providing L2 input and to ensure all participants understood the tasks. For more detail on the writing tasks, see Chapter 3 (Section 3.3.).

5.4.4. Measures of vocabulary knowledge and vocabulary use

The productive vocabulary tests were scored dichotomously, with 1 point for a correct answer and 0 points for an incorrect answer. We adopted a strict scoring system whereby responses were only deemed correct if they were spelled correctly. While this system did not reward learners for partial knowledge of word forms, it ensured that only fully known word forms

were included in the estimate of participants' productive vocabulary knowledge.

In a similar vein, the writing tasks were analyzed exactly as submitted by the participants (albeit anonymized). We operationalized learners' vocabulary use in terms of lexical diversity and lexical sophistication. Diversity was measured by the moving-average type-token ratio (MATTR), which is based on the type-token ratios of consecutive same-size text segments (usually 50 words for written texts). Consequently, it is stable across text lengths (e.g., Zenker & Kyle, 2021). Sophistication was measured as use of low-frequency words by means of both a count-based measure, average log frequency (ALF), as well as a band-based measure, lexical frequency profile (LFP). ALF is the average log-transformed word frequency based on corpus frequency counts, whereas LFP is the proportion of low-frequency words based on 1,000-word bands ordered by decreasing frequency. The commonly used cut-off point of 2K was used to represent low-frequency word use in the LFP measure.

The lexical measures (MATTR, ALF and LFP) were analyzed following Caltabellotta et al. (2025). We used two corpora containing subtitles from various TV shows and movies: Subtlex-US (Brysbaert & New, 2009) and Lexique (New et al., 2007). Learners' output was first lemmatized using spaCy (Honnibal et al., 2020). Next, MATTR was calculated using an existing Python package (Kyle et al., 2021), and ALF and LFP were computed in Python based on corpus frequency counts. We used the lemma frequencies to calculate the average frequency of the learners' texts, and divided the lemmas in 1,000 word frequency bands to calculate the proportion of words beyond the 2K frequency band.

5.4.5. Statistical analyses

The first RQ aimed to investigate the relationship between out-of-school contact, length of instruction, gender and four vocabulary measures (form recall knowledge, lexical diversity, count-based lexical sophistication and band-based lexical sophistication). We used Partial Least Squares Structural Equation Modeling (PLS-SEM) to model these relationships. An important advantage of SEM is that it allows modeling of complex relationships whereby a single variable can be both a dependent and independent variable within the same model. For instance, one can simultaneously specify a path from gender to out-of-school contact (dependent), and from out-of-school contact (independent) to form recall knowledge. In addition, while one must always be cautious about claiming causality based on non-experimental data, the paths in PLS-SEM represent directional relationships, inasmuch as this is theoretically supported (Hair et al., 2021). We used PLS-SEM rather than covariance-based SEM because

it makes no assumptions about data distribution and is therefore robust against non-normality in the data (Hair & Alamer, 2022).

We used the seminr package in R (R Core Team, 2024; Ray et al., 2024) to run separate PLS-SEM models for L2 English and L2 French. First, the time estimates learners provided for each of their out-of-school L2 activities were combined into a composite score (i.e., one score per learner for English and one for French). This allowed us to examine the impact of overall L2 contact on measures of productive vocabulary knowledge and use. Second, length of instruction was calculated based on when learners reported having had their first L2 class (see Learner questionnaires). Third, we also included gender. Since only few participants reported a different gender identity than male (n = 153) or female (n = 230), these participants had to be excluded from the PLS-SEM, bringing the sample size for this analysis to 383 participants. Finally, we included learners' productive vocabulary knowledge and written vocabulary use in the models, operationalized as form recall knowledge (total percentage score), lexical diversity (MATTR), count-based lexical sophistication (average log frequency; ALF) and band-based lexical sophistication (lexical frequency profile; LFP).

In each PLS-SEM model, we entered paths from out-of-school contact, length of instruction and gender to each of the vocabulary measures. These are variables that have previously been shown to be related to receptive vocabulary knowledge (e.g., Peters, Noreillie, et al., 2019; Sundqvist, 2024; Webb & Chang, 2012). Following gender-related findings (e.g., Peters, Noreillie, et al., 2019), we also specified a path from gender to out-of-school contact. This was the simplest, most parsimonious model. Additionally, we also tested a second model including paths from the vocabulary measures to out-of-school contact. This allowed us to explore the directionality of their relationship, as some scholars have suggested that positive relationships could be bidirectional, and they may also reflect that learners with better vocabulary knowledge seek out more out-of-school L2 activities (e.g., De Wilde, 2025). For English, this second model with paths in both directions seemed to perform slightly better based on the lower AIC (for model fit statistics, see the OSF: https://osf.io/bdmwh/). For French, however, the second model performed worse than the first model, based on both the AIC and the BIC. Therefore, we will report the results of the second model for L2 English (including paths from vocabulary measures to out-of-school contact), while we will report the results of the first model for L2 French (excluding these paths).

The second RQ aimed to home in on the relationship between specific out-of-school activities and vocabulary measures. In a first step, we ran exploratory correlation analyses. Due to non-normality, Spearman rank

correlations were computed between the reported time estimates for each of the 11 out-of-school activities (TV without subtitles, TV with L1 subtitles, TV with L2 captions, TV with subtitles in another language, gaming, YouTube, social media, music, reading, speaking and writing to communicate with friends/family and to create content) and the four vocabulary measures (form recall knowledge, MATTR, ALF and LFP). The Bonferroni-adjusted results of the correlational analyses can be found in the Online Supporting Information[30].

In a second step, we ran more complex models to examine the predictive power of out-of-school activities on the vocabulary measures. We conducted a Principal Components Analysis to reduce the number of variables (for the full PCA results, see the Online Supporting Information; OSF: https://osf.io/562nm/). For English, there were three components: *active engagement* = TV with L2 captions, reading, speaking and writing to create content, and to communicate with friends/family; *digital media use* = YouTube, gaming; *passive media exposure* = TV with Dutch subtitles, music, social media. For French, there were two components: *written and audiovisual media* = reading, TV with L2 captions, TV with L1 subtitles, TV without subtitles, social media; *digital media and interaction* = YouTube, TV with subtitles in another language, speaking and writing to create content and to communicate with friends/family, music. We used these principal components to compute Generalized Additive Mixed Models (GAMMs) because there was evidence of non-linearity. The vocabulary measures were modeled as dependent variables and the components were included as smooth terms. A random effect was included for class group to account for the nested data (e.g., VocEN ~ s(PC1) + s(PC2) + s(PC3) + s(ClassGroup, bs = "re")). When the residuals plots showed non-normality, we opted for a non-gaussian distribution if this improved the residuals plots and the model fit (e.g., R^2, *AIC*). Outliers were removed from the French models (one for form recall and LFP, two for MATTR) after visual inspection of the residuals plots. For French, heteroscedasticity was still observed in the residuals plots after these modifications, so we will report bootstrapped standard errors to account for this limitation.

5.5. Results

5.5.1. Descriptive findings

The results of the vocabulary measures and the questionnaire data on out-of-school L2 contact are displayed in Table 5.1. and Table 5.2. The scores on the English vocabulary test indicate that learners were able to

[30] The online supporting information will be made publicly available upon publication of the article. Until then, the information can be accessed via https://osf.io/562nm/.

recall on average 63% of the target items. For French, learners' productive vocabulary knowledge was limited, with an average score of 25%. The mean MATTR was 0.75 for English and 0.69 for French, representing higher lexical diversity in English than French. Average log frequency was observed to be higher for English than French. The mean LFP showed that about 6% of the words in learners' English texts were beyond the 2,000 most frequent words, and 5% of the words in the French texts.

The means for the English out-of-school activities indicate that the participants spent the most time listening to music and using social media (6–8 hours per week on average). Other popular English-language activities were watching TV (without subtitles, with L1 subtitles and L2 captions) and YouTube videos. Overall, learners were on average exposed to 33 weekly hours of out-of-school English, which translates to approximately 4.5 hours per day. For French, however, learners were on average exposed to only 3 weekly hours of out-of-school French. Although all activities had low average time investment, the most common activities included listening to French music, watching TV with L1 subtitles and subtitles in another language, and using social media. A large group of participants (43%) reported being exposed to out-of-school French for less than 1 hour per week, with 23% indicating absolutely no exposure at all, compared to only 1% and 0.5% respectively for out-of-school English. This was supported by the median duration of out-of-school French exposure, which was 0 for all activities except music.

The descriptive findings by gender can be found in the Online Supporting Information (OSF: https://osf.io/562nm/). For English, boys performed better on the form recall test and produced more low-frequent words in their L2 writing. For French, there seemed to be no differences between boys' and girls' vocabulary knowledge and use. In terms of out-of-school L2 contact with English, boys reported higher levels of gaming and YouTube, while girls engaged more frequently in using social media, watching TV with L1 subtitles, speaking and writing to create content, and listening to music. For French, the differences seemed smaller, although a similar trend could be observed: girls watched more French-spoken TV with L1 subtitles and listened to more French music, whereas boys watched more YouTube videos in French.

Table 5.1. Means and standard deviations for learners' form recall knowledge (%), lexical diversity (MATTR), and count-based (ALF) and band-based lexical sophistication (LFP)

	Form recall		**MATTR**		**ALF**		**LFP**	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
English	62.92	17.51	0.75	0.03	3.33	0.09	0.06	0.02
French	24.90	11.79	0.69	0.04	2.62	0.13	0.05	0.02

Table 5.2. Descriptive results for learners' length of instruction (in years), overall out-of-school contact and exposure to specific out-of-school activities, reported in hours per week

	Median	**Mean**	***SD***	**Maximum**[a]
English				
Length of instruction	4.00	3.69	0.96	7.00
Overall out-of-school contact	28.63	33.34	21.16	127.00
TV without subtitles	2.00	3.32	4.45	24.00
TV with L1 subtitles	2.00	3.08	3.87	24.00
TV with L2 captions	1.00	2.93	3.93	23.00
TV with other subtitles	0.00	0.07	0.53	6.00
YouTube videos	2.00	3.30	4.39	21.00
Social media	5.00	6.56	5.70	24.00
Gaming	0.00	2.39	4.33	24.00
Music	6.50	8.32	6.34	24.00
Reading	0.50	1.57	2.66	16.00
Speaking or writing (contact with friends)	0.50	1.26	2.42	20.00
Speaking or writing (creating content)	0.00	0.54	1.88	23.00
French				
Length of instruction	7.00	6.28	1.25	9.00
Overall out-of-school contact	1.00	3.43	7.09	63.00
TV without subtitles	0.00	0.12	0.46	5.00
TV with L1 subtitles	0.00	0.56	1.48	17.00
TV with L2 captions	0.00	0.11	0.66	8.00
TV with other subtitles	0.00	0.40	2.12	23.00

YouTube videos	0.00	0.16	0.89	9.25
Social media	0.00	0.48	1.29	15.00
Gaming	0.00	0.02	0.16	2.00
Music	0.25	1.01	2.16	20.00
Reading	0.00	0.11	0.78	13.00
Speaking or writing (contact with friends)	0.00	0.35	1.22	10.00
Speaking or writing (creating content)	0.00	0.11	0.62	7.00

[a] The minimum value was 0 for all activities, so the maximum also represents the range.

5.5.2. RQ1: The relationship between out-of-school contact, instruction, gender and vocabulary measures

The relationships between out-of-school contact, length of instruction, gender and vocabulary measures were modeled in a PLS-SEM for each L2, visualized in Figure 5.1. and Figure 5.2. The L2 English PLS-SEM (Figure 5.1.) only explained a small amount of variance (R^2 ranging between 0.06 and 0.15). The bootstrapped path coefficients indicated that out-of-school contact was significantly related to all four vocabulary measures, with relatively small effect sizes ($\beta_{\text{form recall}}$ = 0.181, p <.01; β_{MATTR} = 0.141, p <.01; β_{ALF} = -0.128, p <.05; β_{LFP} = 0.127, p <.01). Length of instruction, on the other hand, was not related to all vocabulary measures. Modest to moderate relationships were found for form recall knowledge (β = 0.318, p <.001) and lexical diversity (MATTR; β = 0.202, p <.001). Count-based lexical sophistication (average log frequency; ALF) showed only a weak association (β = -0.105, p <.05), whereas band-based lexical sophistication (lexical frequency profile; LFP) was not related to length of instruction. Gender did not have a direct relationship with total out-of-school contact. However, a modest effect of gender was observed on form recall knowledge (β = -0.154), ALF (β = 0.188) and LFP (β = -0.229) (p <.001), indicating that boys demonstrated better vocabulary knowledge and higher lexical diversity and sophistication. Finally, the PLS-SEM for English also included paths from the vocabulary measures to out-of-school contact. Only the path from form recall knowledge to out-of-school contact was significant (β = 0.170; p <.05), suggesting that learners with better productive vocabulary knowledge tended to engage in more out-of-school English activities.

For L2 French, the PLS-SEM (Figure 5.2.) also explained only a small amount of variance (R^2 between 0.01 and 0.16). The bootstrapped path coefficients indicated moderate significant paths from length of instruction to form recall knowledge (β = 0.386; p <.001) and MATTR (β = 0.303; p <.001). No other paths in the model were significant, indicating no

relationship between of out-of-school French and the vocabulary measures, nor any impact of gender on out-of-school French or vocabulary knowledge and use in French.

Figure 5.1. PLS-SEM results for L2 English

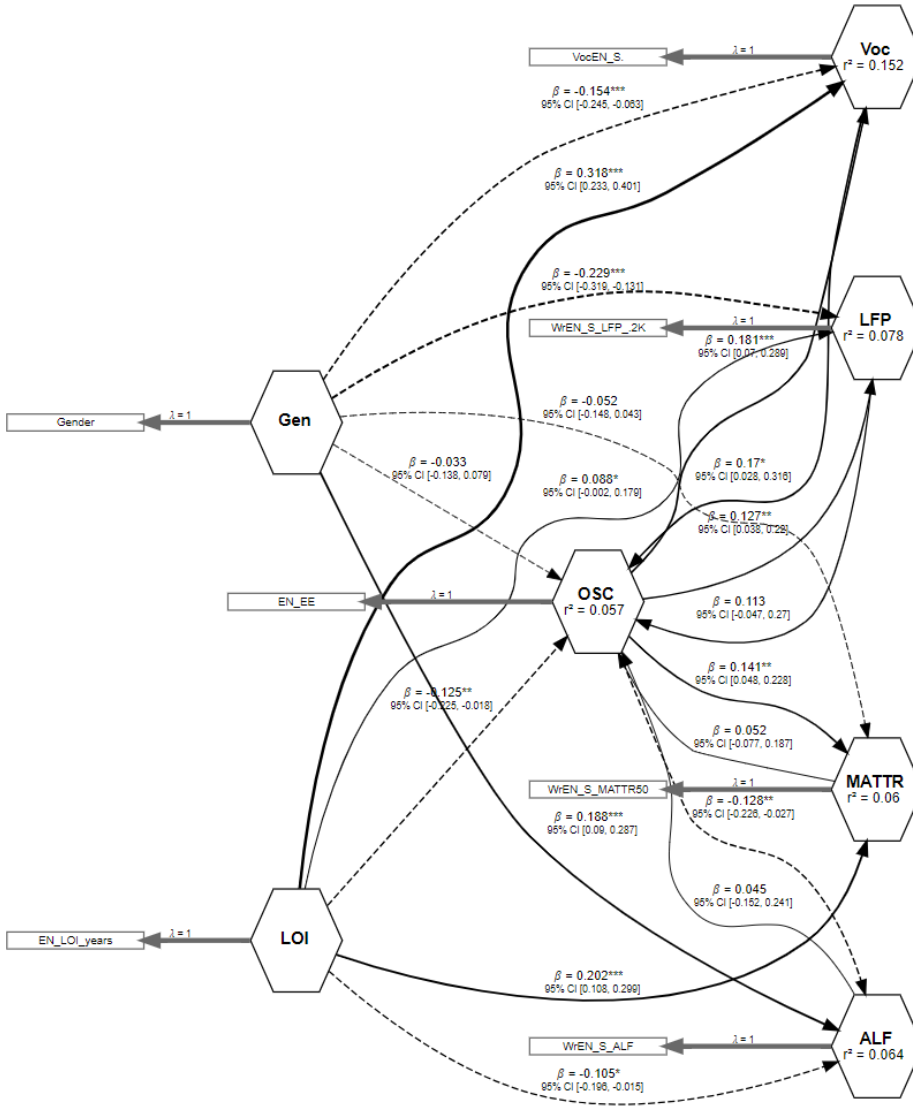
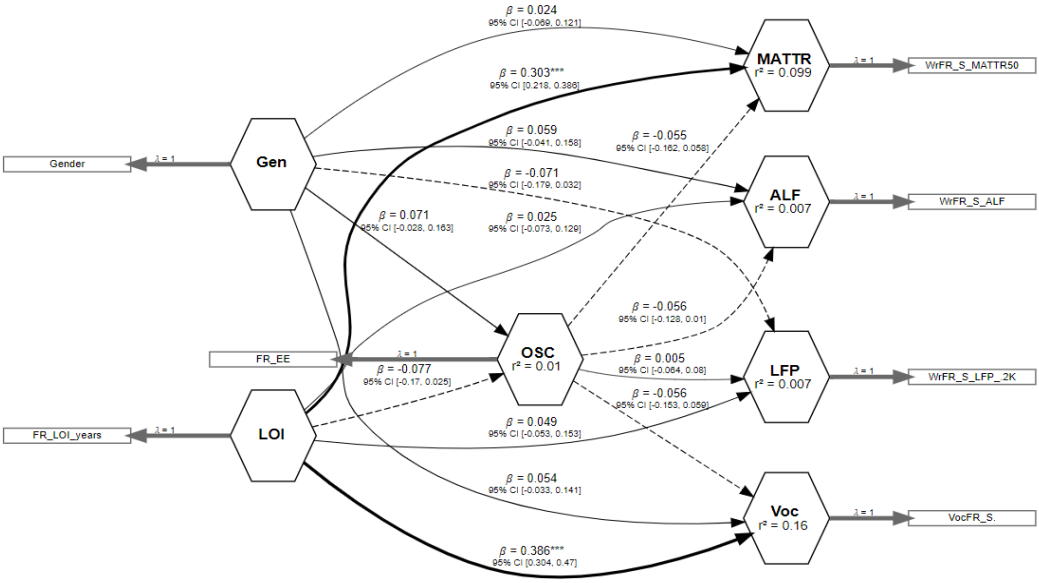


Figure 5.2. PLS-SEM results for L2 French



5.5.3. RQ2: The role of out-of-school L2 activities in vocabulary knowledge and use

In order to investigate the relationship between out-of-school activities and productive vocabulary knowledge and use in more detail, we ran GAMMs to model the predictive power of the principal components (out-of-school contact) on the vocabulary measures (form recall knowledge, MATTR, ALF, LFP). The results tables can be found in the Online Supporting Information (OSF: https://osf.io/562nm/), but findings of GAMMs are best interpreted through plots (shown below). For L2 English (Figure 5.3.), a significant positive impact was found of *active engagement* (TV with L2 captions, reading, speaking and writing) and *digital media use* (YouTube, gaming) on all measures of vocabulary knowledge and use (R^2 form recall = 0.40; MATTR = 0.19; ALF = 0.11; LFP = 0.09). *Passive media exposure* was only significantly related to form recall and LFP, albeit negatively (more out-of-school contact through TV with Dutch subtitles, music and social media predicted lower scores). These findings are largely in line with the exploratory correlational analyses, which showed (a) small positive correlations between form recall knowledge and TV without subtitles, with L2 captions, reading and YouTube; and between lexical sophistication and Youtube and gaming, (b) a small negative correlation between form recall knowledge and TV with L1 subtitles, and (c) no correlations between any of the vocabulary measures and music or social media.

The GAMMs for L2 French (Figure 5.4.) indicated little connection between out-of-school contact and the vocabulary measures, showing no significant relationships. This is in line with the exploratory correlational analyses, which returned no significant correlations after Bonferroni adjustment. The plots show broad confidence intervals beyond the left side of the plots, suggesting limited data for higher ranges of out-of-school contact, which is supported by the descriptive findings for out-of-school French.

Figure 5.3. Plots visualizing the Generalized Additive Mixed Model results for English

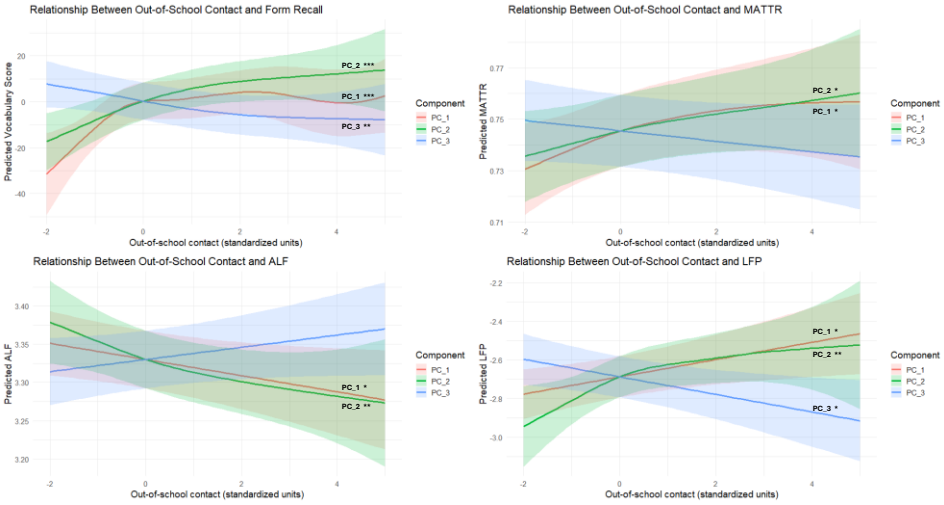
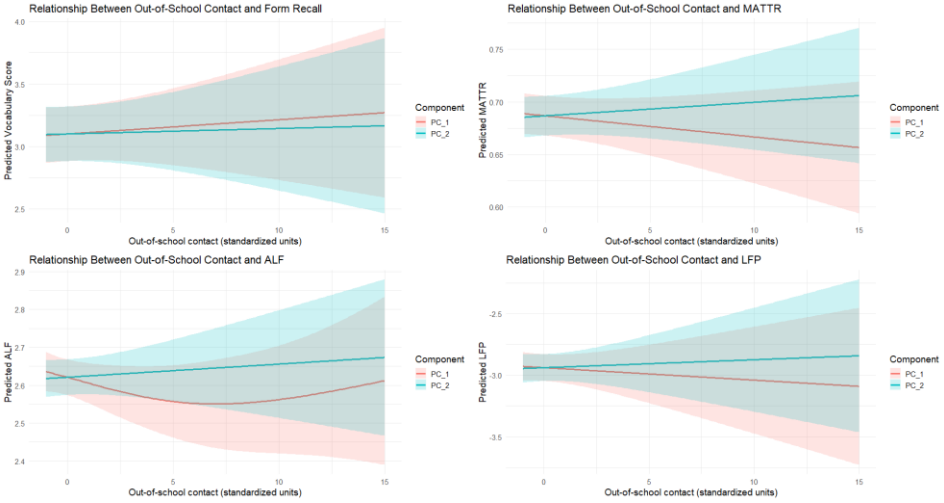


Figure 5.4. Plots visualizing the Generalized Additive Mixed Model results for French



5.6. Discussion

5.6.1. The relationship between out-of-school contact, instruction and vocabulary measures

L2 English

RQ1 aimed to investigate the relationship between out-of-school contact, length of instruction, gender and four vocabulary measures (form recall knowledge, lexical diversity, count-based and band-based lexical sophistication). Learners had extensive out-of-school contact with L2 English (on average 33 hours per week). The PLS-SEM for L2 English showed significant paths from out-of-school contact to all vocabulary measures, which confirms prior research indicating that time spent on out-of-school L2 contact is beneficial to learners' receptive vocabulary knowledge (e.g., Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019), productive vocabulary knowledge (e.g., Bollansée et al., 2021; Sundqvist, 2019), and lexical diversity and sophistication in writing (e.g., Kaatari et al., 2023; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). These findings support usage-based theories of L2 acquisition (e.g., Tomasello, 2003), which emphasize the importance of L2 exposure. While L2 contact in the classroom is limited to a few hours per week, the range of possibilities for out-of-school English allows learners to engage with the L2 more frequently. However, it should be noted that the path coefficients were generally lower than has been found in studies on receptive vocabulary knowledge. It seems likely that the relationship between out-of-school contact and productive vocabulary knowledge and use is less straightforward because it represents a higher level of mastery than receptive knowledge (e.g., González-Fernández & Schmitt, 2020).

In addition, in the English PLS-SEM we also specified a path in the opposite direction, as it seems plausible that more proficient learners might choose to engage with out-of-school English more frequently. Only the path from form recall knowledge to out-of-school contact was significant. This tentatively suggests that the impact of out-of-school contact on vocabulary knowledge and use is more pronounced and straightforward than in the opposite direction. Higher proficiency learners may thus not necessarily choose to engage with more out-of-school English (Peters, Noreillie, et al., 2019), which was also supported by the small path coefficient from length of instruction to out-of-school contact.

Furthermore, the PLS-SEM for L2 English indicated that length of instruction had a positive impact on learners' form recall knowledge, lexical diversity and count-based lexical sophistication, corroborating previous studies on receptive and productive vocabulary knowledge (e.g., Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024; Unsworth

et al., 2014). Band-based lexical sophistication (LFP), however, was not related to length of instruction. This lack of a relationship may be in part due to the nature of the LFP. Given the 1,000-word bands on which the measure is based, the increase in low-frequency word use between grades may be too small to be identified by the LFP (Caltabellotta et al., 2025). Since count-based measures have been shown to better capture differences between proficiency levels (Crossley et al., 2013), this might explain why the average frequency did display a relationship with length of instruction, although the effect size was smaller than for form recall and lexical diversity. The limited improvement in low-frequency vocabulary use with increased instruction may be related to a lack of in-class practice at using low-frequency words (Levitzky-Aviad & Laufer, 2013).

L2 French

The PLS-SEM for L2 French revealed a different picture, showing no significant relationships between total out-of-school French and vocabulary knowledge and use. This is likely because learners had only little exposure to out-of-school French. Since extensive L2 contact may be a prerequisite for out-of-school L2 learning (Ellis & Collins, 2009; Tomasello, 2003), the limited contact with out-of-school French may not suffice to impact on learners' productive vocabulary knowledge and written vocabulary use. In addition to the limited exposure to French, the linguistic distance and morphological differences between French and Dutch may make it more challenging for Dutch-speaking learners to pick up new words in French. For instance, a study by Lu & Dang (2023) with Chinese learners of L2 English showed that L2 contact was not related to learners' receptive vocabulary knowledge, possibly in part due to the linguistic distance between Chinese and English.

Further, significant relationships were established between length of instruction and vocabulary measures, but only for form recall knowledge and MATTR. The effect sizes for length of instruction were higher in the French model than in the English model, which is likely due to the fact that there was limited out-of-school French contact to complement the L2 classroom time. Nonetheless, the model demonstrated no significant paths from length of instruction to lexical sophistication. It is possible that, based on the low form recall scores, the learners' French vocabulary knowledge was still so limited after additional years of instruction that they were not able to produce more low-frequency words (Caltabellotta et al., 2025).

5.6.2. The role of out-of-school activities in vocabulary knowledge and use

L2 English

RQ2 aimed to zoom in on the relationship between out-of-school activities and productive vocabulary knowledge and use. The GAMMs for L2 English showed positive relationships between form recall knowledge, and active engagement and digital media use (TV with L2 captions, reading, speaking and writing; YouTube, gaming). This confirms previous studies showing the benefits of watching TV with L2 captions (Bollansée et al., 2021; Puimège & Peters, 2019b; Pujadas & Muñoz, 2020), as well as watching YouTube videos, gaming, reading, and speaking and writing (e.g., Bollansée et al., 2021; Peters, Noreillie, et al., 2019; Sundqvist, 2019; Sundqvist & Wikström, 2015). In contrast, passive media exposure (TV with L1 subtitles, music, social media) was negatively related to form recall knowledge, lending support to Bollansée et al.'s (2021) findings and contradicting Kuppens (2010). Since this component comprised activities which may not require deliberate attention to the L2, this may suggest that learners do not (or to a lesser extent) actively engage with the English-language input during these activities and, as a result, they may be less beneficial to enhance learning.

For lexical diversity (MATTR), our findings indicated that in general, learners with more out-of-school L2 contact were able to produce more varied vocabulary in their writing. Active engagement and digital media use showed significant positive relationships with lexical diversity. This corroborates findings by Kaatari et al. (2023), who found that having English-language conversations and watching TV positively impacted MATTR (although they found no significant relationship between MATTR and gaming). Similar to form recall, the plot showed a negative relationship between MATTR and passive media exposure, although this finding was non-significant.

Furthermore, we analyzed the relationship between out-of-school contact and lexical sophistication. The results showed that active engagement and digital media use were associated with ALF and LFP, indicating that learners with more L2 contact produced more low-frequency words. Similar to form recall and MATTR, passive media exposure showed an opposite relationship with lexical sophistication than the other components (although non-significant for ALF). While research on the relationship between out-of-school contact and lexical sophistication is scarce, these results seem to confirm previous studies on the role of gaming in written vocabulary use (Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Yet, our findings contradict

those in Olsson and Sylvén (2015), which might be explained by their focus on academic rather than low-frequency vocabulary use.

Finally, it should be noted that the patterns for active engagement and digital media use were very similar. This suggests that while they represent distinct components, their impact is similar, possibly due to their nature. In contrast with passive media exposure, the activities in the other components may require learners to pay attention to the L2 in order to engage in the activity, for instance, in order to follow a storyline (on TV or during reading), play a game, or communicate with friends. This could also explain Sundqvist's (2019) finding that overall time spent on gaming played a more important role in learners' L2 vocabulary knowledge than the level of interaction in games. Indeed, learners may need to master certain vocabulary in order to play parts of the game (e.g., 'pickaxe' in the game *Fortnite*), requiring them to actively use the L2 even if they are not interacting with other players.

L2 French

For French, the GAMMs confirmed the findings of the PLS-SEM in that there was no significant relationship between out-of-school activities and the vocabulary measures. This was not surprising, given that for most activities in the principal components, over 80% of the participants reported that they did not engage in these activities at all (exceptions are TV with L1 subtitles and social media with 60%, and music with 40%). It seems that Flemish learners' out-of-school contact with L2 French is notably lower than with L2 English, as two previous cross-linguistic studies in a similar context have demonstrated (De Wilde et al., 2022; Peters, Noreillie, et al., 2019). This low degree of out-of-school L2 contact with French is likely an important reason for the limited impact of out-of-school contact on vocabulary measures.

5.6.3. Limitations and future directions

There are some limitations to this study. First, our questionnaire items gauging out-of-school L2 contact required participants to accurately estimate the average amount of time per week they spent on each activity. This provided us with estimates that likely approximate learners' total time engagement with out-of-school contact, which allowed us to establish general patterns and trends (see also, e.g., De Wilde et al., 2020b; Kaatari et al., 2023; Lu & Dang, 2023; Sundqvist & Wikström, 2015). It should be noted that the questionnaire did not capture learners' cumulative out-of-school contact, which may be an avenue for future research.

Second, related to this, we inquired about learners' time investment in L2 activities, which means we had no information about learners' level of engagement with the activities. While we posited that watching TV with L1 subtitles, listening to music and using social media do not always require deliberate attention to the L2, it is possible that learners did actively engage with the L2 (e.g., looking up lyrics and singing along). As some researchers have proposed, future research might want to investigate not only behavioral engagement (frequency/duration of the activity), but also dimensions like cognitive engagement (level of attention), affective engagement (enjoyment), and linguistic engagement (focus on linguistic features) (see also Arndt, 2023).

Finally, while we found clear relationships between out-of-school contact and productive vocabulary knowledge and use in L2 English, no significant relationships were observed between out-of-school French and the vocabulary measures. In our sample, this was likely due to the low levels of exposure to L2 French and learners' low proficiency. Future research may benefit from examining learners with higher levels of out-of-school contact or higher proficiency levels, which could shed light on whether out-of-school contact can also benefit L2 vocabulary knowledge and use in L2s other than English, especially when the L1 and L2 are morphologically and etymologically different (see also Lu & Dang, 2023).

5.7. Conclusion

Our study is one of the few studies to investigate out-of-school L2 contact in relation to learners' productive vocabulary knowledge and written lexical diversity and lexical sophistication in two languages, English and French. It operationalized lexical sophistication by means of two indices (count-based + band-based measure) to provide a more nuanced analysis.

The results showed that learners who engaged more frequently with L2 English outside of the classroom performed better on a form recall test. They also produced writing output that was more lexically diverse and contained more low-frequency vocabulary, as computed by both a count-based and band-based measure. A positive relationship was established between all vocabulary measures in L2 English, and active engagement (TV with L2 captions, reading, speaking and writing) and digital media use (YouTube, gaming). However, passive media exposure (TV with L1 subtitles, music and social media) emerged as a negative predictor of two of the vocabulary measures (form recall knowledge and band-based sophistication).

In addition to out-of-school L2 contact with English, length of instruction also benefitted learners' productive vocabulary knowledge, lexical diversity, and count-based lexical sophistication. No relationship was observed with band-based sophistication, which may indicate that count-based measures can better capture small differences between learners (Crossley et al., 2013). It should be noted, however, that increased length of instruction also implies increased length of total L2 exposure, including through out-of-school contact. Given the large amounts of out-of-school English, these two forms of L2 contact cannot be fully distinguished.

For L2 French, a different picture emerged. The findings did not show any relationship between out-of-school L2 contact and productive vocabulary knowledge, lexical diversity and lexical sophistication, likely because out-of-school contact with French was too limited. Length of instruction, however, showed moderate relationships with form recall knowledge and lexical diversity, but not with count-based or band-based lexical sophistication.

In conclusion, our findings underscore the potential benefits of out-of-school L2 contact complementing the L2 classroom for learners' productive vocabulary knowledge, lexical diversity and two measures of lexical sophistication. At the same time, they highlight differences between L2s that likely reflect discrepancies in the amount of out-of-school contact. Taken together, our findings have implications for L2 learning beyond the classroom, suggesting that extensive out-of-school L2 contact is important to effectively support the development of productive vocabulary knowledge and vocabulary use.

6. A methodological perspective on measuring vocabulary use[31]

Abstract

Vocabulary use is an important predictor of L2 writing, and indices of lexical diversity and sophistication are widely used to measure L2 learners' written vocabulary use (Kyle, 2020b; Read, 2000). The present study aims to make a methodological contribution to the assessment of vocabulary use in L2 writing. Although several studies have investigated the reliability of lexical indices (e.g., Crossley et al., 2013; Zenker & Kyle, 2021), few studies have looked at the impact of methodological decisions during pre-processing of learner texts. The present study adopted four operationalizations of lexical indices based on lemmatization (default or custom) and word inclusion (all words or content words). We investigated the effect of these operationalizations on lexical findings in L2 English and L2 French, and their relationship with learners' productive vocabulary knowledge as a proxy for lexical proficiency. The results indicated that both lemmatization and word inclusion had a significant impact on lexical findings in English and French for indices of lexical diversity (moving-average type-token ratio) and lexical sophistication (average log frequency; lexical frequency profile). Overall, default lemmatization of all words appeared to be most strongly related to learners' productive vocabulary knowledge, which suggests that this operationalization best reflected the participants' lexical proficiency.

Keywords: vocabulary use, L2 English, L2 French, lexical diversity, lexical sophistication, L2 writing

[31] Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., & Peters, E. (under review). Measuring vocabulary use in L2 English and L2 French writing: How methodological decisions shape the results.

6.1. Introduction

Vocabulary is considered a key component of writing in a second language (L2). For instance, in large-scale writing assessments like TOEIC, the user scoring guide repeatedly refers to the range and quality of vocabulary in the test takers' essays (Educational Testing Service, n.d.). Vocabulary is also of great interest to L2 research, as it is an important aspect of L2 writing performance. Research has shown that more proficient L2 learners tend to use more varied words (lexical diversity) and more advanced words (lexical sophistication) in their own output (e.g., Consolini & Kyle, 2024; Crossley, 2020; Kisselev et al., 2025; Kyle, 2020b; Treffers-Daller et al., 2018).

In the past decades, there has been a trend in L2 research towards increased methodological rigor (Ortega, 2013a), illustrated by several methodological studies that have investigated the reliability of different lexical indices (e.g., Crossley et al., 2013; Treffers-Daller et al., 2018; Zenker & Kyle, 2021). However, before indices of lexical diversity and lexical sophistication can be computed, researchers go through a pre-processing stage during which they have to determine which counting unit the computations will be based on (word family, flemma, lemma, orthographic type) and which set of words will be included in the calculations (content words and/or function words). While an increasing number of studies are using the lemma (the base form of a word and its inflections) as a counting unit to measure L2 writing in English because it may be more appropriate for productive skills (Kyle, 2020b; Schmitt, 2010a), few studies have investigated morphologically richer languages such as L2 French. Given the complex inflectional system in French, a lower-level lemmatization scheme which reduces the number of word forms under one lemma, may better represent learners' knowledge (see e.g., Consolini & Kyle, 2024, for Spanish). However, this has not been investigated yet. In addition, some recent studies have calculated or reported lexical findings based on content words only (e.g., Díez-Ortega & Kyle, 2024; Guo et al., 2013; Kim et al., 2018). Yet, it is unclear whether excluding function words from the analyses may better represent learners' lexical proficiency than including both content and function words.

The present study aims to make a methodological contribution by shedding light on the impact of methodological decisions during the pre-processing stage with regard to lemmatization and word inclusion. It compares four operationalizations of lexical indices by adopting a default and custom lemmatization scheme, and by computing each of the indices for all words and content words only. The study compares these operationalizations for three indices of vocabulary use: lexical diversity, measured as the moving-average type-token ratio (MATTR), and lexical sophistication, measured by means of both a band-based measure, the

lexical frequency profile (LFP), and a count-based measure, the average log frequency (ALF). The learner corpus consists of written texts in L2 English and L2 French (written by the same learners), which allows an empirical investigation of how lemmatization and word inclusion may affect lexical outcomes in two different L2s.

6.2. Literature review

6.2.1. Vocabulary use in writing

Vocabulary use is an important aspect of L2 learners' writing performance and it has been of increasing importance to researchers. Vocabulary use is often measured in terms of lexical richness (e.g., Kyle, 2020b; Read, 2000), also referred to as lexical complexity (e.g., Bulté & Housen, 2012). Although researchers have adopted slightly different definitions of these constructs, they tend to agree that lexical richness is largely characterized by the type of words learners choose to use. Three well-known and commonly used subconstructs of lexical richness are lexical density, lexical diversity and lexical sophistication (Crossley, 2020; Kyle, 2020b; Read, 2000).

First, lexical density is a measure reflecting how densely information is packaged in a text. It is usually measured as the proportion of content words (as opposed to function words). Since the number of content words can be strongly related to the grammatical and syntactical structures that are used (e.g., *the old wooden chair* versus *the chair which was old and had been made out of wood*), some scholars have argued that this is a measure of syntax rather than vocabulary (Durrant et al., 2022). While it seems to be a good measure to distinguish different registers and modalities, it has not commonly been found to be related to L2 proficiency and may not be the most suitable measure of L2 writing performance (Durrant et al., 2022; Kyle, 2020b). It will therefore not be further discussed in this study.

Second, lexical diversity is characterized by the range or variety of words in a text. The simplest index of lexical diversity is the type-token ratio (TTR): the ratio of unique, unrepeated words in a text to the total number of words (W. Johnson, 1944). The original TTR has been widely criticized for its dependence on text length, and so have mathematical transformations such as Root TTR, Maas' Index and D (e.g., McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). In a comparative study, Zenker and Kyle (2021) found that two measures, the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010) and the moving-average type token ratio (MATTR; Covington & McFall, 2010), were both stable even for texts of under 100 words. MATTR appears to be more strongly related to human

judgements of lexical diversity (Jarvis, 2017), which may indicate that MATTR values represent human ratings of L2 writing better than MTLD. MATTR is computed as the mean TTR based on the TTRs of same-size 'windows' of consecutive, overlapping text segments, usually 50-word windows (e.g., Kyle & Eguchi, 2023; Zenker & Kyle, 2021). Research has shown a strong positive relationship between measures of lexical diversity, including MTLD or MATTR, and L2 proficiency (e.g., Consolini & Kyle, 2024; Engber, 1995; Kisselev et al., 2025; Sung et al., 2024; Treffers-Daller et al., 2018). Since more proficient L2 learners tend to have more productive vocabulary knowledge (Caltabellotta et al., 2025; Laufer, 1998), they have more vocabulary available to avoid repetition in their output, and thus, they can produce more lexically diverse texts. Indeed, some studies have demonstrated that learners with more productive vocabulary knowledge used more varied vocabulary in speaking (Uchihara et al., 2020) and writing (Caltabellotta et al., 2025; Henriksen & Danelund, 2015). Nevertheless, lexical diversity indices do not account for the different types of words that are used, some of which may be perceived as more advanced (e.g., *dog* versus *canine*).

The third subconstruct of lexical richness is lexical sophistication, which is related to the use of sophisticated or advanced words. Although lexical sophistication is increasingly viewed as a multi-dimensional construct (e.g., Kim et al., 2018), its traditional interpretation is related to the use of low-frequency words. More proficient learners tend to use more low-frequency words in their writing (e.g., Crossley, 2020; Kyle et al., 2018; McNamara et al., 2010), and research has also found a relationship between learners' productive vocabulary knowledge and lexical sophistication in speaking (Uchihara et al., 2020) and writing (Caltabellotta et al., 2025; Laufer & Nation, 1995). Some studies have employed the lexical frequency profile (LFP), which can be operationalized as a condensed profile with one value representing the proportion of words beyond the 2,000 most frequent word families (e.g., Laufer, 1998). However, because the LFP depends on 1,000-word frequency bands based on word lists ranked by decreasing frequency, it may not always detect small differences between proficiency levels (Caltabellotta et al., 2025; Crossley et al., 2013). Researchers are now increasingly turning to count-based rather than band-based measures of low-frequency vocabulary use, which use direct frequency counts from a reference corpus instead of frequency bands. The average log frequency, the (log-transformed) mean frequency of the words in a learner text, has been shown to distinguish rather well between proficiency levels (e.g., Crossley et al., 2013; Kyle et al., 2018). Still, Crossley et al. (2013) recommend integrating both count- and band-based indices for a more comprehensive analysis.

6.2.2. Lemmatization

An important aspect of measuring vocabulary use is the counting unit on which the computations are based (e.g., word family, flemma, lemma, orthographic type). Measures like the LFP have traditionally used the word family as a counting unit, implying that all occurrences of a stem, its inflections and its derivations (e.g., *act, actor, acted, active, reactive*) are grouped under one lexical item. This logically affects lexical outcomes because all of these occurrences are counted as the same type, which may impact on the total type count in the learner texts (lexical diversity). Moreover, it can also affect frequency counts (lexical sophistication), because this counting unit does not account for frequency differences between word forms within a word family. Researchers have thus argued that the word family may not be appropriate for productive skills like writing, because it assumes extensive productive knowledge of the inflectional and derivational system (Iwaizumi & Webb, 2022; Kyle, 2020b; Schmitt, 2010a; Schmitt & Zimmerman, 2002). More recent studies have commonly used the lemma or flemma as counting unit (e.g., Kyle et al., 2018), which cluster together the stem and its inflections, but not derivations (e.g., *act, acted, acting*, but not *actor*, *reactive*). Lemmas distinguish between parts of speech (e.g., *act* as a verb and noun under separate lexical items), whereas flemmas do not (*act* as a verb and noun under the same lexical item). A study by Treffers-Daller et al. (2018) that compared different counting units in the analysis of L2 English essays, demonstrated that counts based on lemmas were better predictors of L2 proficiency than counts based on word families. A final operationalization of counting units is at the level of the orthographic type, which means each word form is counted as a separate type (e.g., *act, acted, actor, reactive* are all counted as separate lexical items). Orthographic types assume that L2 learners have no knowledge of inflectional morphology, which may lead to an underestimation of the vocabulary learners can actually use in writing (Kyle, 2020b).

Careful lemmatization seems even more essential in studies involving highly inflected languages, such as French (Treffers-Daller, 2013), because lemmas in morphologically rich languages are comprised of more inflected word forms. Lemmatization can then help to reduce the discrepancies in morphological richness between two languages, although it may not eliminiate them altogether (Treffers-Daller & Korybski, 2015). For instance, while the English verbal lemma *run* has four word forms (*run, runs, running, ran*), the French lemma *courir* has many more inflected forms (*cours, court, courons, courez, courent, courant, couru, courrai, courras, courra,* among others). Consolini and Kyle (2024) investigated L2 learners' vocabulary use in L2 Spanish, another language with an extensive inflectional system. They accounted for morphological richness in Spanish

by adopting a lemmatization scheme in which verbal lemmas were distinguished based on tense and mood, thereby decreasing the number of word forms categorized under one type (see also Díez-Ortega & Kyle, 2024). For French, for example, this would mean that word forms in different tenses are classified as separate lemmas, and learners' knowledge of word forms in the indicative present tense (e.g., *cours, court, courons*) is not assumed to extend to knowledge of the subjunctive present tense (e.g., *coures, coure, courions*) or the indicative past tense (e.g., *courais, courait, couraient*). However, very few studies have taken a similar approach to reduce the number of inflections that are assumed to be known under one lemma, and little research has directly compared multiple lemmatization schemes. As a result, it remains unclear whether a custom lemmatization scheme may actually make a difference in lexical analyses, and whether it would provide a better reflection of learners' productive vocabulary knowledge.

6.2.3. Word inclusion

A second aspect of measuring vocabulary use is the set of words in learner texts that are included in the computations. The most straightforward approach is to compute lexical indices for all words appearing in learner texts. This includes two types of words: content words, open-class words conveying lexical meaning (e.g., *love, ring, devoted*), and function words, closed-class words carrying grammatical meaning (e.g., *le, il, pour*). Function words are essentially grammatical words, like articles, pronouns and prepositions. They are frequent in daily language use (e.g., Dang & Webb, 2016), leading to high frequency counts, and they tend to be regularly repeated, leading to a high token but low type count. Content words, on the other hand, include nouns, lexical verbs, adjectives and adverbs. Copular and auxiliary verbs are usually not considered content words (Engber, 1995; Kyle, 2020b), because they are similar to function words in that they occur frequently and are regularly repeated.

The inclusion of function words in measures of vocabulary use can impact on the outcomes of lexical analyses. The illustrations given in Bottini and Le Foll (2025, p. 52) clearly demonstrate that mean frequency calculations produced highly disparate results when they were based on either all words, content words or function words only. In the past decade, an increasing number of studies have based their computations of lexical indices on content words only, and indices such as content word frequency have indeed shown potential as a predictor of L2 proficiency in English (Bottini & Le Foll, 2025; L. Guo et al., 2013; Kim et al., 2018a) and Spanish (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024). A study by Monteiro et al. (2020) suggested that raw frequency of content words (based on COCA

Academic) was more strongly related to L2 writing proficiency in English than raw frequency of all words. However, the decision to compute lexical indices for content words or all words is not usually justified in the reporting of the methodology, which complicates insights into the impact of this methodological decision, and little research to date has directly compared how the analysis of content words or all words may lead to different lexical outcomes. As a result, it is currently still unclear to what extent this decision regarding word inclusion may truly affect lexical findings, or how strongly these different operationalizations are related to learners' productive vocabulary knowledge.

6.3. Rationale and research questions

Researchers are increasingly aware of the impact of methodological decisions on empirical findings (e.g., Bottini & Le Foll, 2025; Maie et al., 2024). While some studies have compared the reliability of lexical indices measuring vocabulary use (e.g., Treffers-Daller et al., 2018; Zenker & Kyle, 2021), few studies have investigated the impact of different approaches to lemmatization and word inclusion on lexical findings. This may be especially relevant for L2s with a richer morphological and inflectional system like French, since (verbal) lemmas in French tend to be comprised of many different word forms of varying frequency, which are all clustered under one lemma. Moreover, the highly frequent and repetitive nature of function words may also impact strongly on indices of lexical diversity and sophistication. However, it is currently still unclear to what extent these decisions regarding lemmatization and word inclusion may affect lexical findings. Moreover, little is known about how well different operationalizations may represent learners' productive vocabulary knowledge as a proxy for lexical proficiency, an important component of overall L2 proficiency (e.g., Uchihara et al., 2020).

The present study presents a methodological perspective on the measurement of lexical diversity (MATTR), count-based lexical sophistication (ALF) and band-based lexical sophistication (LFP) in L2 writing. To extend the scope of the methodological implications, we investigated written vocabulary use in L2 English and L2 French. In order to investigate the impact of methodological decisions during pre-processing regarding lemmatization and word inclusion, the study adopts four operationalizations of three lexical indices. The following research questions are addressed:

RQ1. To what extent are lexical findings for MATTR, ALF and LFP in L2 English and L2 French affected by four different operationalizations based on lemmatization (default or custom) and word inclusion (all words or content words)?

RQ2. Which operationalization shows the strongest relationship with learners' productive vocabulary knowledge as a proxy for lexical proficiency in L2 English and L2 French?

6.4. Methodology

6.4.1. Design and participants

This study adopted a corpus-based approach in which we analyzed written output by L2 learners of English and French. The L2 texts were taken from a learner corpus collected by Caltabellotta et al. (2025). Participants were 423 learners of English and French in Flanders, Belgium (L1 = Dutch), who each completed two informal writing tasks and a productive vocabulary test in both English and French. The learners were secondary school students in grade 10 (15–16 years old) and grade 12 (17–18 years old), who had had about 3 to 5 years of English instruction and 5 to 7 years of French instruction. Still, they could be considered less proficient in French than English, in large part due to discrepancies in out-of-school L2 contact (Caltabellotta et al., under review; Peters, Noreillie, et al., 2019). We analyzed the learner texts for lexical diversity (MATTR), count-based lexical sophistication (ALF) and band-based lexical sophistication (LFP), each under four operationalizations (see 6.4.3. Lexical analyses). Because the analysis of MATTR required at least 50 tokens per text due to the 50-word window size, the sample size was ultimately reduced to 422 participants for English and 398 participants for French. More detail on the participants can be found in Chapter 3 (Section 3.2.).

6.4.2. Instruments

Writing tasks

The participants completed four informal writing tasks (see Caltabellotta et al., 2025), two in English and two in French, which were adapted from Schoonen (2012) (see also Schoonen et al., 2003). The participants were given two topics per L2, one related to winning a prize and one related to making a complaint, because vocabulary use can be affected by the prompt (e.g., Zenker & Kyle, 2021). The prompts were similar across the two languages to elicit similar output from the participants, but not identical in order to reduce test effects. The writing prompts were provided in the L1 (Dutch) so that learners could not copy any L2 words from the prompt in their own writing. For more detail on the writing tasks, we refer to Chapter 3 (Section 3.3.).

Productive vocabulary test

In addition to the writing tasks, the participants completed a productive vocabulary test in both English and French. The productive VocabLab tests measure form recall knowledge of 60 target items among the 3,000 most frequent words (see also Caltabellotta et al., 2025; Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019). They are fill-in-the-gap tests which require participants to complete sentences with a missing word. The first letter of the word is always provided, as well as the number of missing letters and an L1 (Dutch) translation. For this study, the productive VocabLab test results were used as a measure of learners' productive vocabulary knowledge, which served as a proxy for their lexical proficiency. The productive VocabLab tests are discussed in detail in Chapter 3 (Section 3.3.).

6.4.3. Lexical analyses

Learner corpus

Although the learners wrote two informal texts per L2, we analyzed their writing output for each L2 as a single text to keep the design and the statistical models as parsimonious as possible. We used spaCy (Honnibal et al., 2020) to tokenize and part-of-speech tag the learner texts (models en_core_web_lg and fr_core_news_lg[32]). We ran the lexical analyses on the anonymized learner texts, without applying any changes or corrections (e.g., correcting spelling errors). While this constituted a stricter approach to learners' vocabulary use than if we had corrected minor errors, it allowed us to ensure that words in the learner texts were included in the analyses exactly as they were submitted and, thus, how participants believed they were supposed to be written and used. Non-existent words (tagged as 'X' by spaCy) were excluded from the analyses because they can inflate lexical diversity scores.

Lexical measures

First, we analyzed lexical diversity of the learners' texts using the moving-average type-token ratio (MATTR). This index is based on the original type-token ratio, but it counters an important limitation of the original type-token ratio, namely, its dependence on text length. MATTR has been shown to be particularly stable across text lengths (e.g., Zenker & Kyle, 2021). We computed MATTR values using Kyle et al.'s (2021) lexical

[32] Both models have high average tagging accuracy (100% for tokenization; >97% for part-of-speech tagging). The accuracy rates can be consulted from the spaCy website. For a more detailed accuracy analysis of English spaCy models, see Kyle and Eguchi (2024).

diversity Python package. Since we relied on 50-word windows, a commonly used window size for written texts, the minimum number of tokens in one text required for analysis was 50. However, due to two of the operationalizations excluding certain words from the learner texts (see Operationalizations), some participants' texts were comprised of fewer than 50 tokens. These participants were removed from the dataset, bringing the final sample size to $N = 422$ for English and $N = 398$ for French (initial $N = 423$ for both L2s).

Second, we measured lexical sophistication using both a count-based and band-based measure (average log frequency; lexical frequency profile). These measures were computed based on two reference corpora: Subtlex-US for English (Brysbaert & New, 2009) and Lexique for French (New et al., 2007). They are both subtitle corpora consisting of 51 and 52 million tokens respectively. We tokenized and part-of-speech tagged the reference corpora ourselves using the same spaCy models as for the learner texts, to ensure consistency in the tokenization and part-of-speech tagging process. We then created a frequency list comprised of the lemma frequency counts (normalized per million words), and sorted the list in order of decreasing frequency to create 1,000-word frequency bands. The count-based measure of lexical sophistication consisted of the average log frequency (ALF) of the lemmas in the learner texts, which may be a stronger predictor of proficiency than band-based measures (Crossley et al., 2013). We also used a band-based measure of lexical sophistication, the lexical frequency profile (LFP), as this is a common measure of lexical sophistication in existing research. We computed the LFP for each learner text, represented by the proportion of lemmas beyond the 2,000 most frequent lemmas ('beyond 2K'). The complete Python codes, partly based on Kyle's (2025) Pylats package, are available in the OSF: https://osf.io/ucvm4/.

Operationalizations

The lexical measures were operationalized in four different ways, based on two criteria with two levels each: lemmatization and word inclusion. For lemmatization, we adopted a default and custom lemmatization scheme (see Table 6.1.). The default lemmatization was based on the standard definition of a lemma, that is, the base form of a word and its inflected forms. The custom lemmatization scheme distinguished verbal lemmas not only based on their base form, but also based on mood and tense. This custom lemmatization therefore accounted for the rich verbal morphology in French especially (see also Consolini & Kyle, 2024). Further, we calculated the lexical measures based on word inclusion: the analyses were performed either on all words, or on content words only.

Content words were defined as nouns, lexical verbs (no auxiliaries), adjectives and adverbs (Engber, 1995; Kyle, 2020b; Kyle & Eguchi, 2023). This led to four values per lexical index for each learner text: default - all words, default - content words, custom - all words, and custom - content words.

Table 6.1. Illustration of default and custom lemmatization

Default lemmatization		**Custom lemmatization**	
English: *to see*			
see_VERB:	see, sees, seeing, seen, saw	see_VERB_Inf:	see
		see_VERB_Ind_Pres:	see, sees
		see_VERB_Ind_Past:	saw
		see_VERB_Part_Pres:	seeing
		see_VERB_Part_Past:	seen
French: *voir*			
voir_VERB:	voir, vois, voit, voyons, voyez, voient, voyais, voyait, voyions, voyiez, voyaient, voyant, vu, verrai, verras, verra, verrons, verrez, verront, verrais, verrait, verrions, verriez, verraient, voie, voies, voyions, voyiez, voient, etc.	voir_VERB_Inf:	voir
		voir_VERB_Ind_Pres:	vois, voit, voyons, voyez, voient
		voir_VERB_Ind_Past:	voyais, voyait, voyions, voyiez, voyaient
		voir_VERB_Part_Pres:	voyant
		voir_VERB_Part_Past:	vu
		voir_VERB_Ind_Fut:	verrai, verras, verra, verrons, verrez, verront
		voir_VERB_Ind_Con:	verrais, verrait, verrions, verriez, verraient
		voir_VERB_Sub_Pres:	voie, voies, voyions, voyiez, voient

6.4.4. Statistical analyses

To answer RQ1, linear mixed-effects models were run to compare the lexical values for each operationalization. The dependent variable in each model was the lexical index, that is, lexical diversity (MATTR), count-based lexical sophistication (ALF), or band-based lexical sophistication (LFP), with separate models for English and French. Fixed effects were added for Lemmatization (Default or Custom, reference level Default) and Word Inclusion (All words or Content words, reference level All words), and a random effect was included for Participant to account for individual variation (e.g., lmer(EN_MATTR ~ Lemmatization*WordInclusion + (1|ParticipantID)). While some models showed mild violations of normality at the tails, this was not deemed a reliability concern considering the robustness of linear mixed models to non-normality (see Schielzeth et al., 2020). The assumption of homoscedasticity was not met for multiple models, so for those models, we will report robust standard errors. No issues with multicollinearity were observed.

RQ2 aimed to examine to what extent productive vocabulary knowledge could predict the lexical indices under each operationalization. Since productive vocabulary knowledge was used as a proxy for lexical proficiency, the underlying idea is that the operationalization most reflective of learners' lexical proficiency is the one best predicted by productive vocabulary knowledge (e.g., the model with the highest explained variance, or the best model fit through AIC). We ran generalized additive models (GAMs), which allow modeling of non-linear relationships through so-called smooths, to analyze the predictive value of productive vocabulary knowledge on each of the operationalizations of MATTR, ALF and LFP. For comparability across operationalizations, only a fixed effect for productive vocabulary knowledge was included in the models (no random effect). Since some of the GAMs showed violations of the assumption of normality, we opted for alternative distributions where necessary (scaled t distribution, beta family). The same distribution was used for all operationalizations of one lexical index to ensure comparability of the four models. We removed outliers to improve the residuals plots only when there was clear evidence of outliers based on the QQ plot and the standard deviation of the residuals (SD >4), which was the case for some of the MATTR and LFP models in French. These participants were removed from all four models per index, so that models for each index were based on the exact same dataset. The final models showed no violations of assumptions except heteroscedasticity in some cases, which we account for by reporting robust standard errors.

6.5. Results

6.5.1. Differences in lexical values between operationalizations

The first aim of the study was to investigate to what extent different operationalizations of lexical diversity, count-based and band-based lexical sophistication impacted the values for these lexical indices. The descriptive findings can be found in Table 6.2. and the correlation matrix in Table 6.3. The results of the linear mixed-effects models are presented in Table 6.4.

Table 6.2. Descriptive findings for MATTR, ALF and LFP for all four operationalizations in English and French

	English *M (SD)*			**French** *M (SD)*		
	MATTR	**ALF**	**LFP**	**MATTR**	**ALF**	**LFP**
Default_aw	0.744	3.365	0.054	0.695	3.441	0.067
	(0.03)	(0.09)	(0.02)	(0.05)	(0.08)	(0.02)
Custom_aw	0.759	3.262	0.075	0.707	3.360	0.080
	(0.03)	(0.09)	(0.02)	(0.05)	(0.08)	(0.02)
Default_cw	0.822	2.474	0.119	0.757	2.625	0.131
	(0.05)	(0.13)	(0.04)	(0.07)	(0.18)	(0.05)
Custom_cw	0.850	2.291	0.168	0.779	2.484	0.157
	(0.04)	(0.13)	(0.05)	(0.07)	(0.17)	(0.05)

Note. Default & custom = default or custom lemmatization; aw & cw = analysis of all words or content words only.
Note. Higher MATTR and LFP values indicate higher lexical diversity and sophistication. Lower ALF values indicate higher sophistication.

Table 6.3. Spearman correlation matrix for the four operationalizations

	Default_ aw	**Custom_ aw**	**Default_ cw**	**Custom_ cw**
English				
MATTR				
Default_aw	1			
Custom_aw	0.982	1		
Default_cw	0.732	0.721	1	
Custom_cw	0.717		0.960	1
ALF				
Default_aw	1			

Custom_aw	0.981	1		
Default_cw	0.859	0.851	1	
Custom_cw	0.820	0.851	0.970	1
LFP				
Default_aw	1			
Custom_aw	0.920	1		
Default_cw	0.981	0.894	1	
Custom_cw	0.892	0.980	0.901	1
French				
MATTR				
Default_aw	1			
Custom_aw	0.985	1		
Default_cw	0.680	0.663	1	
Custom_cw	0.697	0.703	0.973	1
ALF				
Default_aw	1			
Custom_aw	0.938	1		
Default_cw	0.769	0.677	1	
Custom_cw	0.752	0.744	0.958	1
LFP				
Default_aw	1			
Custom_aw	0.906	1		
Default_cw	0.978	0.893	1	
Custom_cw	0.866	0.972	0.900	1

Note. All correlations were significant at p <.001.

Moving-average type-token ratio (MATTR)

For lexical diversity, the linear mixed models showed similar results for both L2 English and L2 French (Table 6.4.). Both lemmatization and word inclusion had a significant impact on MATTR values (p <.001), showing that MATTR increased with custom lemmatization and analysis of only content words. Word inclusion showed a larger effect than lemmatization: the analysis of only content words (compared to all words) led to an increase in predicted MATTR values of 0.08 in English and 0.06 in French, while custom lemmatization (compared to default lemmatization) led to a predicted increase of 0.01 in both L2s. The interaction of custom lemmatization and only content words led to a predicted increase of 0.01. The fixed effects (lemmatization and word inclusion) explained 57% of the variance in English and 25% in French. With a random effect for participant, the models accounted for 90% and 83% of the variance respectively, indicating considerable individual variation between participants.

Average log frequency (ALF)

For count-based lexical sophistication (Table 6.4.), a similar trend was observed, with custom lemmatization and content word inclusion leading to lower ALF values (i.e., more low-frequency word use; p <.001). Custom lemmatization led to a predicted decrease in ALF of 0.10 for English and 0.08 for French, while content word analysis resulted in decreased ALF by 0.89 and 0.81. Predicted ALF was reduced by 0.08 (English) and 0.06 (French) when custom lemmatization was applied to only content words. For ALF, individual variation played a less prominent role than for MATTR: the fixed effects alone already explained 95% and 90% of the variance in English and French respectively, with the random effect increasing these numbers to 99% and 97%.

Lexical frequency profile (LFP)

The four operationalizations also significantly impacted the lexical values for band-based lexical sophistication (Table 6.4.), indicating that custom lemmatization and content word inclusion resulted in a larger proportion of low-frequency words (p <.001). With custom lemmatization, the predicted proportion of lemmas beyond 2K increased by 0.02 in English and 0.01 in French. Analysis of only content words resulted in a rise in LFP of 0.07 and 0.06 respectively. When custom lemmatization was combined with only content words, the model predicted an increase in LFP of 0.03 in English and 0.01 in French. The explained variance of the fixed effects was 64% for English and 49% for French, falling in between the R^2 for MATTR and ALF. With the random effect added, the models explained 92% and 91% of the variance.

Table 6.4. Results of linear mixed-effects models for MATTR, ALF and LFP in English and French

	Marginal R^2 / Conditional R^2		**(Intercept)**	**Lemmatization [custom]**	**WordInclusion [cw]**	**Lemmatization × WordInclusion**
English						
MATTR	0.573 / 0.897	Estimate	0.744	0.014	0.077	0.014
		p	<0.001	<0.001	<0.001	<0.001
		CI	[0.741, 0.747]	[0.014, 0.015]	[0.074, 0.080]	[0.013, 0.015]
ALF	0.947 / 0.992	Estimate	3.364	-0.103	-0.891	-0.079
		p	<0.001	<0.001	<0.001	<0.001
		CI	[3.356, 3.373]	[-0.105, -0.101]	[-0.898, -0.885]	[-0.081, -0.078]
		Robust *SE*	(0.004)	(0.001)	(0.003)	(0.001)
LFP	0.640 / 0.919	Estimate	0.054	0.021	0.065	0.027
		p	<0.001	<0.001	<0.001	<0.001
		CI	[0.052, 0.056]	[0.021, 0.022]	[0.063, 0.067]	[0.026, 0.028]
		Robust *SE*	(0.001)	(0.000)	(0.001)	(0.001)
French						
MATTR	0.251 / 0.826	Estimate	0.698	0.012	0.059	0.010
		p	<0.001	<0.001	<0.001	<0.001
		CI	[0.693, 0.702]	[0.011, 0.012]	[0.055, 0.064]	[0.009, 0.011]
		Robust *SE*	(0.002)	(0.000)	(0.002)	(0.001)
ALF	0.903 / 0.969	Estimate	3.443	-0.081	-0.808	-0.062
		p	<0.001	<0.001	<0.001	<0.001
		CI	[3.435, 3.451]	[-0.084, -0.078]	[-0.821, -0.795]	[-0.064, -0.059]
		Robust *SE*	(0.004)	(0.001)	(0.007)	(0.001)
LFP	0.485 / 0.905	Estimate	0.066	0.013	0.063	0.013
		p	<0.001	<0.001	<0.001	<0.001
		CI	[0.063, 0.068]	[0.012, 0.014]	[0.061, 0.066]	[0.012, 0.014]
		Robust *SE*	(0.001)	(0.000)	(0.001)	(0.001)

6.5.2. Relationship between vocabulary knowledge and each operationalization

The second aim of the study was to explore under which operationalization productive vocabulary knowledge was the strongest predictor of the lexical indices. Four GAMs were fitted for each lexical index to explore the predictive power of vocabulary knowledge for each of the operationalizations. We interpreted the models in terms of their R^2 (explained variance in the dependent variable; higher is better) and the AIC (an index of model fit; lower is better).

L2 English

The results of the English GAMs are shown in Table 6.5. For MATTR, the R^2 and AIC showed a comparable trend: custom and default lemmatization of all words showed the highest explained variance (custom_aw = 0.24; default_aw = 0.22) and the best model fit (custom_aw = -1887; default_aw = -1887). Both of these operationalizations outperformed default and custom lemmatization of only content words. We observed the same pattern for ALF, where the explained variance was also highest for custom and default lemmatization of all words (custom_aw = 0.27; default_aw = 0.24), and AIC was the lowest (custom_aw = -921; default_aw = -934). With regard to LFP, the results were less straightforward. While custom lemmatization of content words explained the most variance (R^2 = 0.24), it returned the least favorable model fit of the four operationalizations (AIC = -1527). The model with the best fit, on the other hand, default lemmatization of all words, explained the least amount of variance (AIC = -2298; R^2 = 0.18). Since the model with custom lemmatization of all words explained nearly as much variance as the best-performing model but had a much better model fit (R^2 = 0.24; AIC = -2153), this operationalization seemed to offer the most propitious balance between predictive strength and model fit.

In sum, it seems that for L2 English, the models with custom lemmatization of all words performed the best in terms of the relationship with productive vocabulary knowledge, followed by default lemmatization of all words. When taking into account the balance between explained variance and model fit, the analysis of content words only did not outperform other operationalizations.

Table 6.5. Results of English GAMs with productive vocabulary knowledge as predictor and four operationalizations of lexical indices as dependent variables

	Default_ aw	Custom_ aw	Default_ cw	Custom_ cw
MATTR				
Intercept	0.744 (<.001)	0.759 (<.001)	0.822 (<.001)	0.851 (<.001)
Smooth *p*	<.001	<.001	<.001	<.001
R^2	0.218	0.237	0.149	0.146
AIC	-1887.086	-1887.713	-1467.344	-1505.394
ALF				
Intercept	3.366 (<.001)	3.262 (<.001)	2.473 (<.001)	2.291 (<.001)
Smooth *p*	<.001	<.001	<.001	<.001
R^2	0.238	0.273	0.178	0.198
AIC	-933.841	-921.214	-630.618	-622.356
LFP				
Intercept[a]	0.054 (<.001)	0.075 (<.001)	0.118 (<.001)	0.167 (<.001)
Smooth *p*	<.001	<.001	<.001	<.001
R^2	0.179	0.239	0.184	0.244
AIC	-2297.789	-2153.177	-1667.698	-1526.884

[a] Note that LFP values were transformed using the beta family (logit). For the sake of interpretability, the intercept values have been back-transformed to their original scale.

L2 French

The results of the French GAMs are presented in Table 6.6. For MATTR, the two best-performing models in terms of R^2 and AIC were custom and default lemmatization of all words (R^2 custom_aw = 0.27; default_aw = 0.24; AIC custom_aw = -1514; default_aw = -1523). They returned higher R^2 and lower AIC values than both default and custom lemmatization of only content words, with default lemmatization of content words showing the weakest performance.

With regard to ALF, the findings were mixed. There were only two models out of the four operationalizations in which productive vocabulary knowledge significantly predicted ALF (default lemmatization of content words and custom lemmatization of all words). The model explaining the most variance, default lemmatization of content words, had the worst

model fit (R^2 = 0.02; AIC = -239). Conversely, custom lemmatization of all words showed the best model fit, but also the second lowest R^2 (AIC = -872; R^2 = 0.003). However, although the explained variance was very low in all four models, default lemmatization of content words achieved an R^2 of about 0.02, compared to the other three models which all explained less than 0.5% of the variance (R^2 = 0.002-0.005). As a result, it was deemed the best model out of the four models, despite its limited absolute predictive power. Custom lemmatization of all words was ranked as the second-best model, supported by the significant smooth, the explained variance and AIC. Productive vocabulary knowledge did not predict ALF under custom lemmatization of content words and default lemmatization of all words.

In terms of LFP, default lemmatization of all words was the best-performing model (R^2 = 0.11; AIC = -1872). The weakest model was custom lemmatization of content words, which explained only little variance and had the highest AIC (R^2 = 0.03; AIC = -1322). The results for default lemmatization of content words and custom lemmatization of all words were inconsistent, with default lemmatization outperforming custom lemmatization in terms of R^2, but not in terms of AIC.

In conclusion, the results for French seemed less straightforward than for English. MATTR was best predicted by productive vocabulary knowledge under custom and default lemmatization of all words. For ALF, on the other hand, the best-performing model was default lemmatization of content words, although all four models explained only little variance. Finally, LFP was best predicted by vocabulary knowledge with default lemmatization of all words. It seems that default lemmatization generally performed rather well for French with our sample, and for two out of three lexical indices its performance was best when all words were included.

Table 6.6. Results of French GAMs with productive vocabulary knowledge as predictor and four operationalizations of lexical indices as dependent variables

	Default_ aw	**Custom_ aw**	**Default_ cw**	**Custom_ cw**
MATTR				
Intercept[a]	0.699 (<.001)	0.711 (<.001)	0.759 (<.001)	0.781 (<.001)
Smooth *p*	<.001	<.001	<.001	<.001
R^2	0.242	0.269	0.148	0.180
AIC	-1523.396	-1513.502	-1167.500	-1170.084

ALF				
Intercept	3.444 (<.001)	3.363 (<.001)	2.637 (<.001)	2.493 (<.001)
Smooth *p*	0.655	0.018	0.016	0.453
R^2	0.002	0.003	0.023	0.005
AIC	-868.332	-872.216	-238.665	-267.524
LFP				
Intercept[a]	0.065 (<.001)	0.079 (<.001)	0.129 (<.001)	0.155 (<.001)
Smooth *p*	<.001	<.001	<.001	0.024
R^2	0.105	0.053	0.077	0.031
AIC	-1872.485	-1863.026	-1335.002	-1321.921

[a] Note that MATTR and LFP values were transformed using the beta family (logit). For the sake of interpretability, the intercept values have been back-transformed to their original scale.

6.6. Discussion

6.6.1. The effect of lemmatization and word inclusion

The first RQ aimed to investigate the effect of lemmatization (default or custom) and word inclusion (all words or content words) on lexical values of MATTR, ALF and LFP. The linear mixed-effects models for both L2 English and L2 French showed significant differences between all four operationalizations, with higher levels of lexical diversity and sophistication observed for custom lemmatization and content words analysis. Word inclusion had a stronger effect than lemmatization. This is not surprising, given that the exclusion of function words from the analyses dismisses a large portion of the learner texts. In fact, Nation (2013) states that nearly half of the tokens in English texts tend to be function words, and Lonsdale and Le Bras (2009) claim that 60% of spoken English texts is based on only 50 function words. Needless to say, the exclusion of function words in lexical analyses can affect indices of lexical diversity and sophistication. Given the high frequency count and high rate of repetitiveness of function words (Kyle, 2020b; Nation, 2013), it is not surprising that the exclusion of function words led to higher levels of lexical diversity and sophistication.

Although to a smaller extent, lemmatization was also shown to have a significant effect on the lexical findings in both L2 English and L2 French, which demonstrates the importance of careful lemmatization during pre-processing. Custom lemmatization led to higher levels of lexical diversity and sophistication, which is in line with Treffers-Daller et al.'s (2018) finding that analyses based on orthographic types inflated lexical diversity values compared to lemma-based analyses. It is logical that in our

study, a custom lower-level lemmatization scheme (with more lemmas comprised of fewer forms) would increase lexical diversity and sophistication, as there are more lemmas being counted separately (number of types), and some of those forms are less frequent than others (frequency counts and bands). In this way, custom lemmatization may better account for the differences between learners in terms of their varied and sophisticated vocabulary use.

6.6.2. The relationship with productive vocabulary knowledge

The second RQ aimed to explore the relationship between each of the operationalizations and productive vocabulary knowledge as a proxy for lexical proficiency. Visual inspection of the slopes in the GAM plots indicated similar predictive power of productive vocabulary knowledge across all operationalizations of each lexical index, in both L2 English and French. Based on the explained variance (R^2) and model fit (AIC) for the English models, custom and default lemmatization of all words seemed to return the models that best represented the relationship between the lexical indices and productive vocabulary knowledge. For French, the operationalization returning the best models seemed to be default lemmatization of all words.

Overall, the inclusion of all words (content and function words) in the analyses led to lexical values more strongly related to the learners' productive vocabulary knowledge than content words only, in both L2 English and French. This is in line with a study by Kyle and Eguchi (2023) that showed a stronger relationship between L2 speaking proficiency scores and MATTR for all words than MATTR for content words. However, it contradicts a study on lexical sophistication in L2 writing by Monteiro et al. (2020), who found that frequency measures based on content words better predicted L2 essay scores than frequency measures based on all words. It is possible that Monteiro et al. found content words to be a stronger predictor because their corpus consisted of learners with a wide range of proficiency levels and L1 backgrounds, which may have made it possible for those learners to write more lexically dense texts with a larger variety of (content) words and more low-frequency (content) words. In addition, the informal writing tasks in our study might have elicited more spoken-like output, which could explain the similar results for speaking in Kyle and Eguchi (2023). In the present study, the stronger relationship between productive vocabulary knowledge and the analysis of all words indicates that the inclusion of function words seemed to make a valuable contribution to the measurement of lexical diversity and sophistication (Milton, 2009).

Furthermore, default lemmatization appeared to better reflect learners' productive vocabulary knowledge than custom lemmatization. This was unexpected, since the custom lower-level lemmatization scheme was designed to better account for differences in the number of types and frequency counts, and thus, was expected to better represent learners' lexical proficiency, especially in a morphologically rich L2 like French (Treffers-Daller, 2013). However, it seems that this lower-level lemmatization scheme was not more strongly related to learners' productive vocabulary knowledge. A potential cause could lie in the low French proficiency of the L2 learners (e.g., Caltabellotta et al., 2025), which might limit their productive lexical resources and therefore their use of diverse and sophisticated lemmas. The predictive gains in the models with custom lemmatization may be limited because learners were not yet able to fully 'exploit' the morphological richness of the French language. The participants' low proficiency in French could also help to interpret why the explained variance in the French models was generally lower than in the English models, even if other variables were kept constant (the same learners, lexical indices, and lemmatization schemes). Another explanation for the slightly better performance of default lemmatization could be related to the productive vocabulary test, a form recall test which tested learners' knowledge of target items up to the 3,000 most frequent lemmas and tested only base forms of words (e.g., infinitive verb forms, singular nouns, masculine adjectives). The fact that the vocabulary test did not measure very infrequent words nor learners' knowledge of the inflectional system, may cause a slightly weaker relationship between the custom lemmatization scheme and the vocabulary test scores.

6.6.3. Limitations and future directions

This study has some limitations. First, the learner corpus consisted of informal writing output by L2 learners of English and French, which means the informal genre might have affected learners' vocabulary choices. It should be noted that Caltabellotta et al. (2025) compared the L2 data to output by L1 speakers of English and French in response to the same prompts. They demonstrated that the L1 speakers produced more diverse and sophisticated texts, indicating that the prompts were able to elicit lexically diverse and sophisticated output. Still, future research could analyze texts written in other genres to examine where the conclusions may differ from the findings in this study.

Second, the present study investigated low- to intermediate-proficiency learners, which was addressed as a potential explanation for the finding that default lemmatization of all words was most strongly related to productive vocabulary knowledge, especially in French. It would

be useful to investigate L2 learners at a higher proficiency level to see whether the same findings would emerge. It is possible that higher-proficiency learners' productive vocabulary knowledge might show a stronger relationship with custom lemmatization and inclusion of only content words because they have more diverse and sophisticated lemmas at their disposal.

Lastly, we analyzed the relationship between the lexical indices and learners' productive vocabulary knowledge, using the vocabulary test scores as a proxy for lexical proficiency. Another interesting approach could be to analyze the relationship between the lexical indices and holistic ratings of the L2 texts. This could provide interesting insights into which operationalization best predicts learners' L2 writing performance as rated by human judgements, since humans' natural sensitivity to variety and frequency of vocabulary may be more strongly related to a custom lemmatization scheme (see also Jarvis, 2017).

6.7. Conclusion

This study aimed to investigate the impact of methodological choices on the measurement of vocabulary use through lexical diversity (MATTR), count-based sophistication (ALF), and band-based sophistication (LFP). It explored the effect of decisions during the pre-processing stage regarding lemmatization and word inclusion, by analyzing L2 texts in English and French under four different operationalizations for each lexical index (default and custom lemmatization; inclusion of all words and content words only). The analyses were performed on a learner corpus comprised of texts written by L2 learners who each completed two writing tasks in both L2 English and L2 French.

The findings indicated that lexical values for MATTR, ALF and LFP all differed significantly across the four operationalizations, with custom lemmatization and content word analysis leading to higher values in both L2 English and French. The relationship with productive vocabulary knowledge was strongest overall for default lemmatization of all words. This may be an artifact of the participants' proficiency level: their vocabulary use may not have been varied and sophisticated enough for custom lemmatization and content word analysis to better reflect their productive vocabulary knowledge (although custom lemmatization performed well for English, the L2 for which learners' productive vocabulary knowledge was better).

Taken together, these findings have important methodological implications for the analysis of lexical diversity and lexical sophistication

in L2 writing output. They clearly demonstrate that methodological choices during the pre-processing stage significantly affect the lexical findings, which is likely to impact on the conclusions drawn. Our results highlight the importance of consistency in lemmatization and word inclusion choices across lexical indices or learner groups being compared, because comparisons may be highly skewed if different procedures are applied. Finally, the findings also seem to underscore the importance of aligning methodological choices with the learner group under investigation (e.g., Kyle, 2020b; Schmitt, 2010a).

7. Conclusions

This PhD thesis brought together three empirical studies on L2 learners' productive vocabulary knowledge and vocabulary use in writing in L2 English and L2 French. In this chapter, I will first recapitulate and summarize the main aims and findings of this PhD project (7.1.). Next, I will discuss how this PhD thesis contributes to our understanding of productive vocabulary knowledge and vocabulary use in L2 English and L2 French by addressing the theoretical (7.2.), pedagogical (7.3.) and methodological (7.4.) implications of the findings. Finally, I will discuss the limitations of the studies and suggest potential avenues for future research (7.5.).

7.1. Summary of the three studies

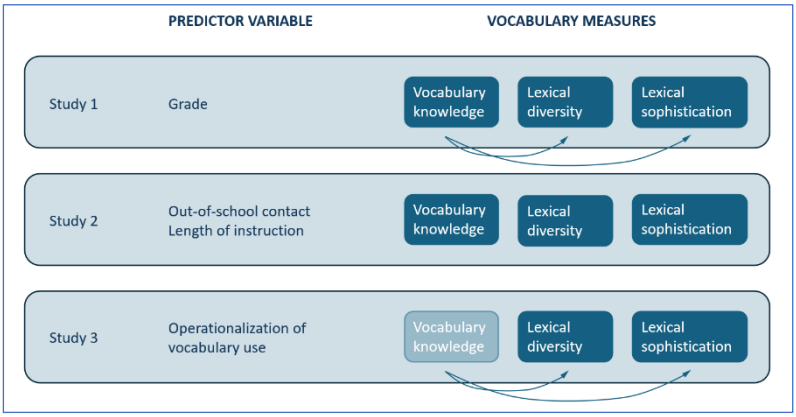
This PhD thesis is the first to present a multi-dimensional investigation of L2 learners' productive vocabulary knowledge and vocabulary use in L2 English and L2 French. First, it set out to investigate the cross-sectional differences in productive vocabulary knowledge (form recall) and written vocabulary use (lexical diversity and lexical sophistication) between two grades, as well as the relationship between productive vocabulary knowledge and written vocabulary use. Second, it aimed to examine the role of formal instruction and out-of-school contact in productive vocabulary knowledge and written vocabulary use. Third, the thesis sought to investigate the impact of methodological decisions on findings of vocabulary use. Through three studies (see Figure 7.1.), the thesis aimed to answer the following general research questions:

RQ1. How do productive vocabulary knowledge (form recall knowledge) and written vocabulary use (lexical diversity and lexical sophistication) develop cross-sectionally in L2 English and L2 French, and how are they interrelated? (Study 1)

RQ2. What is the role of out-of-school L2 contact and length of formal instruction in productive vocabulary knowledge, lexical diversity and lexical sophistication in written L2 English and L2 French? (Study 2)

RQ3. What is the impact of different operationalizations of lexical diversity and lexical sophistication on the outcomes of lexical analyses in L2 English and L2 French, and how are they related to productive vocabulary knowledge? (Study 3)

Figure 7.1. Overview of the three studies



The thesis contributes to existing L2 research by focusing on a large sample of secondary school learners, two L2s (English and French), and measures of productive vocabulary knowledge and vocabulary use, rather than the more commonly investigated receptive vocabulary knowledge. The multi-dimensional nature of the thesis becomes apparent through several key features of the studies. First, the focus is on two aspects of productive mastery of vocabulary knowledge: knowledge of the form-meaning link at the form recall level, and free vocabulary use in writing. Second, the thesis draws on two constructs to examine learners' vocabulary use (lexical diversity and lexical sophistication), supporting the need for multivariate approaches to vocabulary use (e.g., Consolini & Kyle, 2024). Third, lexical sophistication was measured through both mean-based and band-based indices (average log frequency and lexical frequency profile), providing a more comprehensive picture of learners' lexical sophistication than single-index analyses (e.g., Crossley et al., 2013). Finally, the thesis investigated vocabulary knowledge and use from multiple perspectives, including both learner-related variables like grade, length of instruction and out-of-school contact, as well as methodology-related factors like the operationalization of lexical indices.

Data were collected from 423 Dutch-speaking L2 learners of English and French in grade 10 (15–16 years old) and grade 12 (17–18 years old), who each completed a test battery in both L2s. For each L2, the participants completed a form recall vocabulary test (productive vocabulary knowledge), two writing tasks (vocabulary use), and a questionnaire measuring their out-of-school L2 contact and length of formal instruction. The vocabulary test measured form recall knowledge of words among the 3,000 most frequent lemmas. The writing tasks were analyzed in terms of lexical diversity (moving-average type-token ratio)

and two indices of lexical sophistication (average log frequency and lexical frequency profile).

7.1.1. Study 1

Study 1 (Chapter 4) aimed to investigate the role of grade in productive vocabulary knowledge, lexical diversity and lexical sophistication by investigating the cross-sectional differences between learners in grade 10 (ages 15–16) and grade 12 (ages 17–18). A second aim was to investigate the relationship between productive vocabulary knowledge and lexical diversity and sophistication in this cross-sectional sample.

The findings of Study 1 demonstrated that on average, Flemish learners of L2 English and L2 French did not master the 2,000 most frequent lemmas at the level of form recall in either L2. For the productive vocabulary test in L2 English, the grade 10 learners obtained average test scores of 66% for the 1–2K level and 45% for the 3K level, compared to average scores of 78% and 58% in grade 12. Following Schmitt et al.'s (2001) suggestion to interpret a score of 26/30 (86%) on each frequency level as mastery of that level, neither grade 10 nor grade 12 learners appeared to have mastered the 2,000 most frequent words productively, although the grade 12 learners came fairly close. A small portion of the participants in grade 10 (12%) and grade 12 (35%) demonstrated mastery of the 2K level in English. Form recall knowledge in L2 French, on the other hand, was much more limited, with grade 10 learners obtaining low average scores of 25% (1–2K) and 13% (3K), and grade 12 learners obtaining scores of 39% (1–2K) and 21% (3K). Learners in neither grade had mastered any of the frequency levels productively, although there was a clear improvement between the grades. Not a single participant in grade 10 had mastered the 2K level in French as measured by our form recall test, and only one participant in grade 12.

In terms of cross-sectional differences in productive vocabulary knowledge and vocabulary use, the results showed that grade 12 learners had better productive vocabulary knowledge and used more varied vocabulary in their writing than grade 10 learners in both L2 English and L2 French. The clear cross-sectional differences between the grades indicate (cross-sectional) development of learners' form recall knowledge and lexical diversity in writing. In contrast, smaller cross-sectional differences were found between the grades in terms of the use of sophisticated or advanced words (viz., low-frequency words). Lexical sophistication was observed to differ between the grades only when using a measure of average word frequency (mean-based sophistication) in English and only to a small extent. The finding that lexical sophistication

did not seem to improve much across grades in either L2 may indicate that lexical sophistication develops more slowly than other aspects of vocabulary knowledge and use. It should be noted that the two lexical sophistication indices provided a different picture of the cross-sectional differences in English, as the small improvement in mean-based sophistication (average frequency) was not reflected in the results for band-based sophistication (lexical frequency profile, i.e., the proportion of low-frequency words). This may be related to the nature of the band-based measure, which may be less sensitive to small advancements in sophistication.

Furthermore, the results of Study 1 demonstrated that the relationship between productive vocabulary knowledge and lexical diversity and sophistication was different in L2 English and L2 French. While productive vocabulary knowledge was related to lexical diversity in both L2s, it was found to be related to lexical sophistication indices only in English. It is likely that this cross-linguistic difference stems from the participants' limited productive vocabulary knowledge in French, which may have led to limited use of more sophisticated words by the learners, but also to difficulty in detecting any potential relationship through statistical models.

7.1.2. Study 2

Study 2 (Chapter 5) extended the first study by investigating how productive vocabulary knowledge, lexical diversity and lexical sophistication may be impacted by learner variables such as out-of-school L2 contact and length of formal instruction. Along with the impact of total out-of-school contact (average overall time investment per week in hours and minutes), the study also examined the role of time spent on specific out-of-school activities.

The findings demonstrated that out-of-school contact with English was extensive (M = 33 hours per week; Mdn = 29 hours per week), and participants reported spending the most time on average on music (M = 8 hours; Mdn = 7 hours), social media (M = 7 hours; Mdn = 5 hours), TV without subtitles (M = 3 hours; Mdn = 2 hours), YouTube videos (M = 3 hours; Mdn = 2 hours), and TV with L1 subtitles (M = 3 hours; Mdn = 2 hours). Overall out-of-school contact with English was found to be related to productive vocabulary knowledge, lexical diversity, mean-based and band-based lexical sophistication. Especially activities that may require more cognitive engagement with the L2 (e.g., watching TV with L2 captions, gaming, reading; as opposed to watching TV with L1 subtitles, listening to music) seemed more strongly related to the vocabulary measures. Out-of-school contact with French, on the other hand, was very

limited (*M* = 3 hours per week; *Mdn* = 1 hour per week) and did not show any relationship with productive vocabulary knowledge, lexical diversity or lexical sophistication.

Nevertheless, in spite of the extensive out-of-school contact for English, length of instruction was still a stronger predictor of English productive vocabulary knowledge, lexical diversity and mean-based lexical sophistication than out-of-school contact. For French, length of instruction seemed to be the main source of L2 input, and it was the only predictor of productive vocabulary knowledge and lexical diversity (but not lexical sophistication). The findings of Study 2 point to a positive impact of formal instruction, while also highlighting the potential of extensive L2 contact for the development of productive vocabulary knowledge, lexical diversity and lexical sophistication. Yet, based on the cross-sectional differences in Study 1, it seems that this may be a fairly slow process, particularly in terms of lexical sophistication.

7.1.3. Study 3

Study 3 (Chapter 6) built on the lexical indices used in the previous studies by taking a methodological approach to the analysis of lexical diversity, mean-based and band-based lexical sophistication. It sought to explore the effect of different methodological decisions during the pre-processing stage on lexical findings. The first methodological decision was lemmatization, which determines which word forms are considered as a single lexical unit. One lemma in English is comprised of fewer word forms than one lemma in French. Given the morphological richness of French, a custom lemmatization scheme which reduces the set of word forms under one lemma, may better represent learners' vocabulary knowledge. The second methodological decision was word inclusion, which reflects whether all words in a text are analyzed or content words only. In light of the frequent and repetitive nature of function words, this decision may have a strong impact on indices of lexical diversity and sophistication. Study 3 adopted four operationalizations of the lexical indices based on lemmatization (default or custom) and word inclusion (all words or content words only). In order to determine the most appropriate operationalization, the study also examined the relationship between each of the operationalizations and learners' productive vocabulary knowledge.

The results indicated similar patterns for L2 English and French, with both lemmatization and word inclusion significantly affecting all lexical indices. Custom lemmatization and the inclusion of content words only led to higher observed levels of lexical diversity and sophistication. However, no single operationalization clearly emerged as most strongly related to productive vocabulary knowledge, although an analysis of the

explained variance and model fit seemed to indicate that the models with all words performed better than those with content words only. Default lemmatization outperformed custom lemmatization for French, whereas default and custom lemmatization both performed fairly well for English. The results could be an artifact of the proficiency level of the participants, which was higher for English than for French. Since the custom lemmatization scheme was designed to better capture differences in repetition and frequency of word forms within a default lemma, it is possible that participants were able to better exploit and use more diverse and sophisticated inflected forms in English than in French. With regard to word inclusion, both L2s led to similar results, with the analyses of all words generally outperforming the analyses of content words only. This indicates that, at least for our sample, function words contributed valuable information to the analyses of the three lexical indices.

7.2. Theoretical contribution

The findings of this PhD thesis contribute theoretically to research on L2 vocabulary learning and development in two L2s, English and French. The thesis sheds light on L2 learners' productive vocabulary knowledge at the form recall level, as well as their lexical diversity and lexical sophistication in writing. More specifically, the thesis informs L2 vocabulary research through its findings with regard to (a) the cross-sectional differences in productive vocabulary knowledge, lexical diversity and lexical sophistication between grade 10 and grade 12 learners; (b) the relationship between productive vocabulary knowledge on the one hand and lexical diversity and sophistication on the other; (c) the role of out-of-school L2 contact and length of formal instruction in productive vocabulary knowledge, lexical diversity and lexical sophistication; and (d) the differential results for L2 English and L2 French within the same learners.

7.2.1. Cross-sectional development

Vocabulary knowledge

This PhD thesis set out to investigate how L2 learners' productive vocabulary knowledge and vocabulary use in writing develops cross-sectionally for two L2s, English and French. With regard to vocabulary knowledge, the results of the form recall tests clearly indicate that the development of form recall knowledge is not straightforward. On average, the L2 learners in neither grade 10 nor grade 12 demonstrated mastery of the 2,000 most frequent lemmas on our form recall tests, and this held true across both L2 English and L2 French. This may indicate that form recall knowledge is an aspect of vocabulary knowledge that is more difficult to

acquire (e.g., Laufer, 1998; Laufer & Goldstein, 2004), in support of González-Fernández and Schmitt's (2020) finding that recall knowledge was consistently acquired after recognition knowledge. The difficulty of increasing form recall knowledge may be due to the fact that fairly precise and unassisted word knowledge is required in order to recall the correct word form based on a meaningful context (Henriksen, 1999; Melka, 1997; Perfetti & Hart, 2002; Schmitt, 2019). The necessary quality and strength of learners' lexical representations may not always be sufficiently developed for learners to recall and produce word forms (fluently) (e.g., Ellis & Wulff, 2020; Perfetti & Hart, 2002). As González-Fernández and Schmitt (2020) point out, it seems to be a complex and time-consuming process to develop vocabulary knowledge 'towards the productive end of the receptive/productive continuum' (p. 501).

Although direct comparisons with previous studies are not evident due to methodological differences, the findings of this PhD project suggest that productive vocabulary knowledge likely develops at a slower rate than receptive vocabulary knowledge. The difference in form recall knowledge between grades implied lower cross-sectional gains than have generally been found for meaning recognition knowledge (Henriksen & Danelund, 2015; Laufer, 1998; Peters, Noreillie, et al., 2019). In a direct comparison of meaning recognition and form recall knowledge, Laufer (1998) showed that the difference between grade 10 and 11 participants was greater for meaning recognition (19% at the 2K level, 32% at 3K) than for form recall (17% at 2K and 3K). In Flanders, Peters, Noreillie, et al. (2019) observed cross-sectional differences in meaning recognition of about 15 to 25% between grade 8, grade 10 and first-year university students in L2 English and French, compared to 11% (French) and 13% (English) in this PhD. The findings of this PhD thesis seem to be in support of Schmitt's (2019) claim that it is more challenging to enhance vocabulary knowledge to productive mastery than to learn words at the receptive level.

This finding also aligns with previous research demonstrating that productive vocabulary knowledge is generally more limited than receptive vocabulary knowledge, with learners performing better on meaning recognition than on form recall tests (e.g., Fan, 2000; González-Fernández & Schmitt, 2020; Henriksen & Danelund, 2015; Laufer, 1998; Laufer & Goldstein, 2004; Noreillie et al., 2020). A comparison with the Flemish grade 10 learners in Peters, Noreillie, et al. (2019) points to a similar pattern, since their participants obtained an average of 87% at the 2K level on an English meaning recognition test and 73% at the 3K level, compared to 66% and 45% on the form recall test in the present PhD thesis. Other studies using meaning recognition tests have also found average scores above 70% at the 2K level for L2 English learners in secondary school (e.g., Laufer, 1998; Noreillie et al., 2020) and at university (e.g., Fan, 2000; Lin, 2023;

Tschirner, 2004; Zhang & Lu, 2014), while test scores on form recall tests have tended to be lower (e.g., Henriksen & Danelund, 2015; Laufer, 1998).

It should be noted that for L2 French, the participants demonstrated even more limited productive vocabulary knowledge, with lower average scores than have previously been found for English (e.g., Henriksen & Danelund, 2015; Laufer, 1998; Webb, 2008) or for French (e.g., Noreillie et al., 2020). In fact, the grade 12 learners in this PhD thesis obtained lower French form recall scores than the Flemish grade 12 learners in Noreillie et al. (2020), even though both studies collected data from L2 French participants in grade 12 in Flanders using the productive VocabLab test. Although the smaller sample size in Noreillie et al.'s study ($N = 50$) could be a contributing factor, the exact reason for the difference in test scores remains unclear. Nevertheless, the low scores on the form recall test are a testament to the participants' limited productive vocabulary knowledge in French, and they seem somewhat in line with (though still lower than) scores by grade 10 learners of L2 English in Israel, who receive most of their L2 English input from formal instruction (Laufer, 1998).

Nevertheless, even though the development of form recall knowledge appears to be a relatively slow and complex process, the findings showed that productive vocabulary knowledge developed cross-sectionally from grade 10 to grade 12 in both L2 English and L2 French. The differences in test scores are largely comparable to those found in Laufer (1998), although Laufer's participants were only one grade apart. About one-third of the Flemish participants had mastered the 2K level in English by grade 12, which is similar to the vocabulary knowledge of Danish grade 12 learners of English observed in Henriksen and Danelund (2015). Taken together, the findings of this PhD project indicate that productive vocabulary knowledge differs cross-sectionally, though presumably to a lesser degree than receptive vocabulary knowledge.

Vocabulary use

With regard to vocabulary use, the findings of this PhD thesis contribute to SLA research by shedding light on L2 learner differences in terms of lexical diversity and lexical sophistication across grades in secondary school. The findings illustrated clear cross-sectional differences in lexical diversity between grade 10 and grade 12 learners. Based on MATTR, grade 12 learners used significantly more varied vocabulary in their writing than grade 10 learners in both L2 English and L2 French, similar to Henriksen and Danelund's (2015) finding that Guiraud's Index values for English L2 texts differed between grades 10 and 12. However, the effect sizes for the cross-sectional differences in lexical diversity were smaller than those for productive vocabulary knowledge in both English and French.

Although we did not statistically determine an 'order of acquisition' of the components of vocabulary knowledge and use (see the implicational scaling method in González-Fernández & Schmitt, 2020), these results may indicate that the ability to use varied vocabulary in writing takes longer to develop than form recall knowledge of the form-meaning link.

In contrast with lexical diversity, lexical sophistication seemed to differ only to a limited extent between grade 10 and grade 12 learners. For L2 English, lexical sophistication was observed to increase with a small effect size when it was measured using a mean-based approach (average log frequency), indicating that grade 12 learners used slightly more low-frequency words. However, lexical sophistication was not found to increase using the band-based approach (lexical frequency profile). For L2 French, lexical sophistication did not differ cross-sectionally between grades with either measure. This limited (or lack of a) difference in lexical sophistication is in line with previous cross-sectional studies with secondary school students (Henriksen & Danelund, 2015; Laufer, 1998; Levitzky-Aviad & Laufer, 2013). These results suggest that lexical sophistication is an aspect of vocabulary which tends to develop at a slower pace. In light of Nation's (2001, 2013) framework on the components of vocabulary knowledge, which distinguishes knowledge of form, meaning and use, this may imply that use is a more difficult aspect of vocabulary knowledge to learn, especially so for less frequent words.

A potential explanation for the limited cross-sectional difference in lexical sophistication between grade 10 and grade 12 may lie in the ease (or difficulty) of lexical access to less frequent word forms. From a functionalist perspective (e.g., Bardovi-Harlig, 2020), writing tasks require learners to retrieve an L2 word form from memory that denotes a concept or meaning they want to convey. This retrieval can be more or less efficient depending on the strength of the lexical representations, which determines how well learners can access the L2 word forms (e.g., Ellis & Wulff, 2020; Puimège, 2022). Models of lexical access (e.g., Forster, 1976; Morton, 1969) postulate that more frequent words, which have higher probability of occurrence, are activated or reached in the mental lexicon before less frequent words. In a writing task, learners may therefore choose a higher-frequency word that roughly represents their intended meaning without considering other, less frequent options, because the higher-frequency word is accessed more quickly and easily. The notion of lexical access highlights the difficulty for L2 learners of using less frequent words in their written output on a writing task[33].

[33] For a brief discussion of how the prompt and task instructions might also impact learners' use of more or less sophisticated words, see Section 7.4.2.

Nevertheless, this pattern may be different for L1 writers. The L1 data analyzed in Study 1 showed that L1 writers of English and French produced more diverse and more sophisticated words in response to the same prompts, which indicates that they retrieved a wider range of words and less frequent words from their mental lexicon. This discrepancy between L2 and L1 writers is likely a consequence of L1 writers' larger vocabulary size (Chenu & Jisa, 2009; Iwaizumi & Webb, 2022; Laufer, 1995; Meara, 1988; Miralpeix, 2020; Nation, 2001; Qian & Lin, 2020), which implies that L1 writers have more words available in their mental lexicon, among which less frequent words. Moreover, due to L1 writers' extensive experience with the language, they have likely developed much stronger lexical representations of these less frequent words (Ellis & Wulff, 2020; Libben, 2022; Pellicer-Sánchez et al., 2022). For L1 writers, less frequent items may be activated in similar ways or at a similar time as more frequent items. The higher-frequency items may then be by-passed or ignored, because the lower-frequency items are equally strongly represented or the lower-frequency words allow for a more specific or nuanced expression of the intended meaning.

7.2.2. The relationship between productive vocabulary knowledge and vocabulary use

This PhD thesis also sought to investigate the relationship between productive vocabulary knowledge and written vocabulary use in L2 English and L2 French, since little is known about how productive vocabulary knowledge relates to free vocabulary use in writing. The results indicated that lexical diversity was strongly related to productive vocabulary knowledge in both L2s. Even for L2 French, where learners' vocabulary knowledge was limited, lexical diversity showed a strong relationship with productive vocabulary test scores, supporting the idea that lexical diversity may already start to develop at an early stage of L2 proficiency. The findings provide support for the notion that lexical diversity is a by-product of vocabulary breadth (e.g., Bulté & Housen, 2012; Díez-Ortega & Kyle, 2024; Olinghouse & Wilson, 2013). The broader a learners' lexical repertoire (i.e., the wider the range of words they know, or simply put, the larger their vocabulary knowledge), the more opportunities they have to avoid repetition in their writing. This is in line with research that has established a clear relationship between proficiency level and the use of varied words (e.g., Consolini & Kyle, 2024; Engber, 1995; Treffers-Daller et al., 2018). Lexical diversity may be an aspect of vocabulary use that develops somewhat naturally in light of humans' natural urge to strive for a balance between variety and repetition, as claimed by Zipf (1935, as cited in Jarvis, 2017). It seems plausible, then, that humans are likely to integrate varied vocabulary in their own output inasmuch as they are able to do so, which

could explain why productive vocabulary knowledge was more strongly related to lexical diversity than to lexical sophistication in Study 1 and Study 3 (see also Henriksen & Danelund, 2015).

At the same time, the findings of this PhD thesis showed a less clear and consistent relationship between productive vocabulary knowledge and lexical sophistication. Taken together with the limited cross-sectional differences in lexical sophistication, this could suggest that the use of low-frequency words develops at a slower pace than the use of varied words or form recall knowledge of the form-meaning link. While productive vocabulary knowledge showed a moderate relationship with the average log frequency in English, the explained variance was lower than for lexical diversity, and there was only a small-sized relationship with the lexical frequency profile. In French, no significant relationships were found between productive vocabulary knowledge and either measure of lexical sophistication. This implies that even if learners have more low-frequency vocabulary available, this does not always mean they will use it in their own writing. Indeed, some other studies have found weak or non-significant relationships between productive vocabulary knowledge and lexical sophistication (Henriksen & Danelund, 2015; Laufer, 1998; Lemmouh, 2010), illustrating that learners with more knowledge of low-frequent vocabulary do not necessarily produce more low-frequent words in their writing. These findings highlight the challenges regarding vocabulary that L2 learners face during writing (Schmitt, 2020).

One explanation for this finding was previously discussed (Section 7.2.1.) and is related to lexical access. Learners with more vocabulary knowledge may not necessarily use more low-frequency words because they can less efficiently access these less frequent words (Conklin, 2020; Ellis & Wulff, 2020; Laufer, 2005a; Libben, 2022; Puimège, 2024). In addition, it has been suggested that L2 learners may adopt a risk-avoiding strategy in a testing context because they are aware they might be penalized for making errors (e.g., Laufer, 1998). This means that even if L2 learners have (partial) knowledge of less frequent words, they may refrain from using them if they are uncertain about how the words should be used (e.g., Fan, 2009; Laufer, 1998; Levitzky-Aviad & Laufer, 2013).

7.2.3. The role of out-of-school contact and formal instruction

Another aim of this PhD thesis was to extend previous research into the impact of out-of-school contact and length of formal instruction on receptive vocabulary knowledge, by investigating their impact on productive vocabulary knowledge, lexical diversity and lexical sophistication. The findings of the empirical studies clearly point to the beneficial role of overall L2 contact in the development of productive

vocabulary knowledge, lexical diversity and lexical sophistication. The findings from Study 1 (differences between grades in formal instruction) and Study 2 (relationships from length of instruction and out-of-school contact to several vocabulary measures, see below) align well with usage-based perspectives on L2 learning, which emphasize the importance of L2 input for language learning (e.g., Ellis & Wulff, 2020; Tomasello, 2003). According to usage-based theories, languages are learned as a result of repeated exposure and use. For vocabulary, this implies that L2 learners first form initial lexical representations of words they encounter in the L2, inside or outside the L2 classroom. Through frequent and repeated exposure to these L2 words, their lexical representations become stronger, facilitating retrieval of word information during word recognition as well as word production (e.g., Ellis & Wulff, 2020; Libben, 2022). For instance, learners who watch the Netflix show *Selling Sunset* might repeatedly encounter words like *estate, property, invest,* and *mortgage*. Indeed, Rodgers and Webb (2011) found that words are more likely to be repeated within a single season of a TV show than across TV shows. With repeated exposure, learners' lexical representations become stronger and can be retrieved more easily and more efficiently during cued recall and free writing, which can be beneficial to learners' performance on a form recall test and their lexical diversity and sophistication in writing tasks (e.g., using the word *property* instead of *building*). This is reminiscent of Krashen's (1982) input hypothesis, in that learners exposed to more L2 input (through instruction and out-of-school contact) tended to have higher levels of productive vocabulary knowledge, lexical diversity and lexical sophistication.

Nonetheless, it should be noted that Krashen's (1982) concept of comprehensible input is limited to L2 input only (no output or interaction) and meaning-focused activities only (no focus on form). However, the results of Study 2 seem to indicate that it is L2 *contact* rather than only L2 *input* which is related to L2 learning. The analysis of different out-of-school activities showed that activities like watching TV with L2 captions, reading, speaking and writing, and gaming were most strongly related to the measures of productive vocabulary knowledge and use. The fact that these activities are not all input-based only, seems to be in support of the criticism of the input hypothesis that the notion of comprehensible input may be too restrictive. The findings of this PhD project point to the usefulness of a broader interpretation of L2 'input' including both input- and output-based exposure to the L2, as has long been the case in usage-based theories (e.g., Wulff, 2021). Instead of the activities being input- or output-based, what could be more important is the level of cognitive engagement involved in the activity (Arndt, 2019; González-Fernández, 2021; Kerekes, 2015; Schmitt, 2008). For instance, learners likely have to actively engage with L2 audio or writing in order to follow a storyline in a novel or in a TV show with L2 captions, and to successfully play a game or

interact with other players across the world. In contrast, activities like watching TV with L1 subtitles and listening to music can realistically be completed with little cognitive engagement (e.g., reading the L1 subtitles may suffice to understand what is happening; music can play in the background while learners are involved in another activity[34]).

Furthermore, the findings of this PhD thesis, and specifically Study 2, highlight the potential of out-of-school L2 contact in the development of productive vocabulary knowledge, lexical diversity and lexical sophistication. While previous research has extensively documented a positive relationship between out-of-school contact with English and receptive vocabulary knowledge (e.g., De Wilde, 2025; Ghamarian-Krenn, 2023; Gyllstad et al., 2025; Peters, Noreillie, et al., 2019; Sundqvist, 2019; Warnby, 2022), this PhD thesis demonstrated that out-of-school contact with English was also positively associated with productive vocabulary knowledge at the form recall level (Bollansée et al., 2021; Sundqvist, 2019; Sylvén & Sundqvist, 2012) as well as the use of lexically diverse (Kaatari et al., 2023; Sundqvist, 2019) and lexically sophisticated words in writing (Sundqvist & Wikström, 2015; but in contrast with Cobb & Horst, 2011; Olsson & Sylvén, 2015). An important nuance based on the findings is that out-of-school L2 contact may need to be extensive before this relationship can be established. It seems that when out-of-school contact is extensive, like in L2 English, learners who are exposed to the L2 more often also demonstrate better vocabulary knowledge and vocabulary use. Yet, no such relationship was found for L2 French, indicating that the limited L2 exposure may not suffice for a relationship with productive vocabulary knowledge and use. This corresponds well with scholars who have emphasized the importance of large quantities of L2 contact by suggesting that vocabulary development can be promoted through extensive reading (e.g., Day & Robb, 2015; Krashen, 2004; Nation & Waring, 2024) and extensive TV viewing (e.g., Pujadas & Muñoz, 2019; Webb, 2015).

Apart from out-of-school contact, the results of Study 1 and Study 2 highlight the value of formal instruction for the development of productive vocabulary knowledge, lexical diversity and lexical sophistication. Study 2 found length of formal instruction to be the strongest predictor of productive vocabulary knowledge, lexical diversity and lexical sophistication in L2 English, and the only predictor of productive vocabulary knowledge and lexical diversity in L2 French. These findings

[34] This by no means denies the possibility that learners do actively engage with the music they listen to, for instance, by singing along, paying attention to pronunciation or looking up the lyrics (Kerekes, 2015). This difference in engagement may be an explaining factor for why some studies have found positive relationships between music and L2 proficiency (e.g., Puimège & Peters, 2019a), whereas others have not (e.g., Hannibal Jensen, 2017; Peters, Noreillie, et al., 2019).

run counter to some previous studies which have suggested that out-of-school contact was a stronger predictor of receptive vocabulary knowledge than formal instruction (Busby, 2021; Peters, 2018; Peters, Noreillie, et al., 2019; Schwarz & Ghamarian-Krenn, 2024; Warnby, 2022). This could mean that formal instruction plays a comparably bigger role in the development of productive knowledge than that of receptive knowledge. The present PhD thesis therefore suggests that the role of out-of-school contact and formal instruction may differ depending on the aspect of vocabulary knowledge. A reason for this may be related to the nature of vocabulary learning in instructional settings, often aimed at deliberate learning. In-class L2 activities may provide focused and repeated practice opportunities with vocabulary in isolation (e.g., flashcards) as well as in communicative use (e.g., writing tasks), through a combination of meaning-focused and language-focused L2 activities (Nation, 2007). The fairly controlled nature of formal instruction, where teachers can decide which words to focus on in which ways and with which frequency, can help learners to expand their knowledge at the productive level in a targeted manner, including form recall knowledge and the ability to use vocabulary in their own output (e.g., Laufer, 2020; Laufer & Paribakht, 1998).

7.2.4. Cross-linguistic differences

The findings of the three empirical studies clearly showed that results could not always be generalized across the two L2s under investigation. This demonstrates the value of cross-linguistic studies investigating multiple L2s, since findings for one L2 may not always extend to other L2s, even within the same L2 learners (e.g., Peters, Noreillie, et al., 2019; Treffers-Daller et al., 2008). In this PhD project, learners' vocabulary knowledge was found to be much more limited in French than in English. While this is in line with a few studies in a Flemish context (De Wilde et al., 2022; Noreillie et al., 2018; Peters, Noreillie, et al., 2019), the cross-linguistic difference in form recall knowledge seems to be larger than has previously been found for meaning recognition knowledge. The smallest cross-linguistic difference found in Peters, Noreillie, et al. (2019) was for the university students at the 2K level, where the participants had a higher score for English than French by 13%, and the largest cross-linguistic difference was a 30% higher score for English than French (grade 10 students at the 3K level). In Study 1 of this PhD thesis, however, the smallest cross-linguistic difference at the form recall level was found to be a difference of 31% (grade 10 at the 3K level), while the largest cross-linguistic difference was a difference of 40% (grade 10 and 12 at the 2K level). In combination with the generally low vocabulary test scores for French, it appears to be more difficult for learners to make progress in their form recall knowledge in L2 French than in L2 English.

Based on the findings of Study 2, it seems likely that out-of-school contact plays an important part in the cross-linguistic differences. Even though L2 learners in Flanders have had two or three years of additional formal instruction in French, this does not seem to compensate for the extensive out-of-school L2 contact with English (De Clercq & Simoens, 2015; De Wilde et al., 2022; Peters, Noreillie, et al., 2019). Flemish learners' out-of-school contact with French, on the other hand, is very limited, which may cause it to have only a limited (statistically non-significant) impact on learners' productive vocabulary knowledge and use. Moreover, research has shown that vocabulary gains from incidental learning tend to be smaller than gains from intentional learning (Laufer, 2003, 2005a; Schmitt, 2008; Webb et al., 2023). Given the often informal, incidental and leisure-based nature of out-of-school activities (Lyrigkou, 2023; Rød & Calafato, 2025; Schwarz & Ghamarian-Krenn, 2024; Sundqvist, 2009), it may be even more difficult for L2 French learners to develop their productive vocabulary knowledge and vocabulary use incidentally when the L2 contact is so limited.

Another reason for the cross-linguistic difference between L2 English and L2 French may be found in the linguistic distance from the learners' L1, Dutch. Linguistic distance is a predictor of L2 performance in that L2s with a larger linguistic distance from the learners' L1 tend to be more difficult to learn (e.g., Berthele et al., 2022; Granfeldt et al., 2023). Since Dutch has a larger linguistic distance from French (a Romance language) than from English (both Germanic languages), L2 French is likely harder to learn for L1 speakers of Dutch than L2 English. The situation in Flanders can be compared to that in the German-speaking cantons (also a Germanic language) in Switzerland, where L2 English and L2 French are both mandatory subjects in school. Some of the German-speaking cantons start formal French instruction in grade 3 and formal English instruction in grade 5, and other cantons introduce the L2s in the opposite order. Although direct comparative studies of L2 English and L2 French are limited, Swiss L2 learners appear to achieve better vocabulary knowledge and higher proficiency in L2 English, regardless of the start of instruction (Berthele, personal communication, June 28, 2025; Berthele & Udry, 2021; Pfenninger & Singleton, 2016, 2019). While this may be partly explained by higher levels of out-of-school contact and motivation for English (Berthele & Udry, 2021; Pfenninger, 2020), Berthele et al. (2022) indicated that linguistic distance also had a small effect on L2 French performance in the four skills (listening, reading, speaking, writing).

7.3. Pedagogical contribution

The present PhD thesis makes a pedagogical contribution to SLA and vocabulary research by demonstrating the role of formal instruction in

the development of productive vocabulary knowledge, lexical diversity and lexical sophistication in writing. More specifically, the results have implications for L2 pedagogy in terms of (a) the selection of vocabulary; (b) the role of in-class and out-of-school activities; and (c) the cross-linguistic differences between L2 English and L2 French, particularly in view of the current attainment targets.

7.3.1. Selection of vocabulary

The results of this PhD project indicated that L2 learners of English and French did not master the 2,000 most frequent lemmas at the form recall level in L2 English or L2 French, with especially limited vocabulary knowledge in French. A possible explanation for this lies in the selection of vocabulary addressed during formal instruction, for instance, through textbooks. Scholars have suggested following a frequency principle in the selection of vocabulary for instruction (He & Godfroid, 2019; Nation, 2013; N. Schmitt & Schmitt, 2014; Vilkaitė-Lozdienė & Schmitt, 2020), with a focus on high-frequency words for beginner learners and gradually more attention to words of lower frequency for more advanced learners. High-frequency words are considered important because they are necessary to function in the L2 (Schmitt & Schmitt, 2014; Webb & Chang, 2012), and they are useful because L2 learners are likely to frequently encounter or use them (He & Godfroid, 2019; Laufer & Nation, 2012). The proposed cut-off for high-frequency words was originally set to 2K (Laufer & Nation, 1995; Nation, 2001), and this cut-off is still commonly used to measure vocabulary use in learners' output. Yet, from a pedagogical perspective, a 3K threshold seems to be more widely adopted nowadays (Nation & Anthony, 2013; Schmitt & Schmitt, 2014; Vilkaitė-Lozdienė & Schmitt, 2020), because the 3,000 most frequent word families in English generally provide around 80–90% lexical coverage (viz., the proportion of words known in a text), allowing learners to understand a large portion of English input (Laufer, 2005a; Nation, 2001; Schmitt & Schmitt, 2014).

However, although high-frequency words are necessary and useful, they may not always be sufficient for authentic L2 encounters, such as reading, listening or watching TV for pleasure (Milton, 2022; Nation, 2006; Schmitt & Schmitt, 2014). While the 2,000 to 3,000 most frequent word families have been suggested for 95% coverage in listening (van Zeeland & Schmitt, 2013) and TV viewing (Webb & Rodgers, 2009), an optimal level of 98% coverage for comfortable, unassisted listening would translate to 6,000 or 7,000 word families (Nation, 2006). The proposed requirements for reading are higher, with 4,000 to 5,000 word families for 95% coverage (Qian & Lin, 2020), and 8,000 to 9,000 word families for 98% coverage (Nation, 2006). Research has therefore suggested that as learners progress

beyond the beginner stage of L2 learning, mid-frequency words (3K–9K) can be integrated, because these are useful in facilitating authentic L2 activities (Nation, 2006)[35]. Against this background, the finding that Flemish learners did not master the 2K level productively, underscores the importance of attention to high-frequency vocabulary before expanding the focus to less frequent words.

Research has shown that textbooks often do not take a systematic or research-driven approach to the selection of vocabulary, as vocabulary is often selected as a function of the topic of reading passages (e.g., Schmitt, 2019). In contrast with the recommended gradual decrease in frequency levels, textbooks often prioritize high-frequency words that learners are likely to already know (Bergström, 2024; Hajiyeva, 2015; Matsuoka & Hirsh, 2010; Sun & Dang, 2020), and they do not always show a consistent decrease in frequency levels across coursebook chapters (Sun & Dang, 2020), secondary school grades (Van Parys et al., 2024), or proficiency levels (Hsu, 2009). Evidence from Flemish textbooks suggests that this issue may be of particular relevance for French: textbooks at the start of secondary school (grade 7) seem to cover only half of the 1,000 most frequent flemmas in French, compared to 60–65% in English (Van Parys et al., 2024). The participants' limited French productive vocabulary knowledge (even at the 2K level) may thus partly stem from unsystematic vocabulary selection in instruction.

Another consequence of the lack of a systematic approach to vocabulary selection is limited recycling of vocabulary across tasks and classes (Schmitt, 2019), going against the well-documented benefits of (spaced) repetition and practice to solidify learners' knowledge (Ellis & Collins, 2009; Loewen, 2020; Nation, 2007; Nation & Webb, 2011; Schmitt, 2019; Webb et al., 2023). Studies have shown that textbooks tend to offer limited opportunities for repeated exposure to vocabulary (Bergström et al., 2025; Matsuoka & Hirsh, 2010; Nguyen, 2021; Van Parys et al., 2024). A study of Flemish textbooks indicated that almost 50% of vocabulary in English and French textbooks was not recycled at all (Van Parys et al., 2024), and a study by Matsuoka and Hirsh (2010) on an ELT coursebook showed that 66% of the words beyond 2K occurred only once. Such findings may partly explain why the participants in this PhD thesis had not mastered vocabulary up to 3K at the form recall level.

[35] Low-frequency words beyond 9K may be less appropriate in a context with limited L2 classroom time, because their low frequency leads them to be less useful for day-to-day L2 communication purposes (He & Godfroid, 2019; Schmitt & Schmitt, 2014). Beyond the 9K level, the relative gains in lexical coverage tend to be only small.

7.3.2. Vocabulary activities

This PhD thesis showed that although learners in grade 12 had more form recall knowledge of 2K and 3K vocabulary than those in grade 10, they did not necessarily produce more words beyond 2K in their own writing. This discrepancy between the words learners can recall and the words they actually use in writing, may be linked to the type of vocabulary activities that are included in formal instruction. One of Nation's (2007) strands, language-focused learning, relates to explicit teaching of vocabulary through activities that require deliberate attention to the words, for instance, through flashcards, sentence writing, or fill-in-the-blanks. Indeed, research has shown that word-focused instruction can be effective in promoting vocabulary learning (Laufer, 2020; Loewen, 2020; Webb et al., 2020), and learning gains from deliberate, language-focused activities tend to be larger than those from incidental, meaning-focused activities (Laufer, 2003; Schmitt, 2008; Webb et al., 2023). Moreover, productive (output-based) activities seem to lead to higher learning gains than receptive (input-based) activities (e.g., Hennebry et al., 2017; Laufer & Rozovski-Roitblat, 2015; Webb, 2009). For instance, Laufer and Rozovski-Roitblat (2015) found that words were more likely to be retained after 2–3 exposures in vocabulary exercises than after 18–21 exposures in a reading text. Productive exercises such as L1-to-L2 translations, sentence writing, and composition writing may be especially conducive to learning, as word production tends to require more engagement with the vocabulary (Griffin, 1992; Laufer, 2003; Laufer & Hulstijn, 2001; Nation & Webb, 2011; Schmitt, 2008).

However, the question arises why the benefits of additional formal instruction in grade 12 were reflected in learners' form recall knowledge, but less so in their use of sophisticated vocabulary, particularly in French. A study of formal instruction in Flanders showed that the opportunities for writing and speaking are comparable in both English and French instruction[36] (Housen et al., 2001), in keeping with Nation's (2007) strand of meaning-focused output (see also Swain, 1985) and the field of task-based language teaching (e.g., Ellis, 2017; Van den Branden, 2022). In spite of these communicative opportunities in class, learners did not use more

[36] There were few differences between English and French in terms of communicative opportunities included in the classes, and the range of speaking and writing opportunities was more dependent on the individual teachers than on the L2. This is in contrast with the general perception of French and English instruction in Flanders, where French instruction is often assumed to be more form- and word-focused (explicit instruction of grammar and vocabulary) than communication-based (Housen et al., 2001, p. 136). It should be noted that this study was conducted in the early 2000s and may no longer fully reflect current teaching practices.

vocabulary beyond 2K even if their form recall scores indicated learning gains at the 3K level.

A potential explanation is that learners adopt a risk-avoiding strategy because they are aware that reduced accuracy of their vocabulary use may affect their grades (e.g., Laufer, 1998). They may refrain from using less frequent words they are uncertain about, and instead choose to rely on high-frequency words they are more familiar with (Laufer, 1998; Levitzky-Aviad & Laufer, 2013; Liao & Fukuya, 2004). In that sense, there may be a trade-off between learners' accuracy and complexity (Bulté & Housen, 2014). In order to improve learners' familiarity with lower-frequency words, teachers could offer targeted exercises (e.g., sentence writing) that encourage deliberate and repeated practice with more advanced and less well-known words. This could help to consolidate learners' knowledge and extend their mastery beyond knowledge of the form-meaning link (e.g., Hunt & Beglar, 1998; Schmitt, 2019). Moreover, to counter learners' fear of being penalized for errors in their output, teachers could offer a safe space that allows learners to experiment with lower-frequency and advanced words in writing, thus prioritizing enhancing complexity first and accuracy later (Bulté & Housen, 2014). This seems to be in favor of formative assessment strategies, which allow learners to receive low-stakes feedback as they practice and make mistakes (Manchón, 2011; Morris et al., 2021).

Another important finding of this PhD thesis was that out-of-school contact can serve as a valuable source of L2 contact for the development of productive vocabulary knowledge, lexical diversity and lexical sophistication (at least for English, where out-of-school contact was extensive). While out-of-school contact is characterized by learner-initiated activities and teachers do not have direct control over this, they can encourage out-of-school L2 contact to increase learners' engagement with the L2[37]. Indeed, Study 2 showed that the time spent on English out-of-school activities was a predictor of productive vocabulary knowledge, lexical diversity and lexical sophistication (e.g., Bollansée et al., 2021; Kaatari et al., 2023; Sundqvist, 2019). The findings of Study 2 suggested that activities requiring more cognitive engagement were more strongly related to productive vocabulary knowledge and use (Arndt, 2023; Schmidt, 1990; Schmitt, 2008), so teachers may especially recommend activities like reading, watching TV with L2 captions, or speaking and writing in the L2. Activities that may typically be more passive or less cognitively engaging, like listening to music or watching TV with L1 subtitles, can also be useful for L2 learning (Kuppens, 2010; Niitemaa,

[37] It is useful to note that out-of-school contact is learner-initiated, often departing from intrinsic motivation to engage in certain activities. As a result, it may be important to let learners retain their autonomy and agency in deciding whether or not to engage with the L2 outside the L2 classroom (Sockett, 2022; Sundqvist, 2009).

2020; Puimège & Peters, 2019a), presumably on the condition that learners are actively involved in the activity (Arndt, 2023; González-Fernández, 2021; Kerekes, 2015; Schmitt, 2008; Sockett & Toffoli, 2012).

In addition, teachers may also bridge the gap between out-of-school and in-school L2 contact by marrying common out-of-school activities to the classroom (Rød & Calafato, 2025; Schmitt, 2019; Sockett, 2025). Out-of-school activities are learner-initiated and, therefore, often accompanied by learners' motivation to engage in the activities for pleasure and enjoyment (Alonso Alonso, 2022; Lyrigkou, 2023; Sundqvist & Sylvén, 2014). Given the positive impact of motivation and enjoyment on L2 learning (Dewaele & Macintyre, 2014; Dörnyei, 2005; Tahmouresi & Papi, 2021), it could be useful to integrate popular out-of-school activities into the classroom, for instance, by including music, TV shows or games into in-class activities. For instance, two small-scale studies have found that L2 English learners had higher vocabulary learning gains when they played *The Sims* while being exposed to supplementary vocabulary exercises, and they considered the supplementary materials enjoyable and helpful for playing the game (Miller & Hegelheimer, 2006; Ranalli, 2008).

7.3.3. Cross-linguistic differences

This PhD thesis underscored notable differences in productive vocabulary knowledge and out-of-school contact between L2 English and L2 French. What could be useful in instructional settings is to try to tailor in-class activities to the learner group and their learning context. Study 2 pointed to a distinct learning context for L2 English and L2 French even though the L2s are both compulsory subjects in secondary school, because learners come into contact with L2 English and French to a highly different extent and in very different ways. Teachers could try to take into account learners' prior knowledge and out-of-school contact. For instance, if learners tend to have receptive knowledge of many high-frequent words from out-of-school contact, as seems to be the case for L2 English in many countries (e.g., Peters, Noreillie, et al., 2019; Sundqvist, 2019; Warnby, 2022), it may be wise to focus classroom time on the development of productive aspects of vocabulary knowledge or on lower-frequency words less prominent in typical out-of-school activities (Olsson & Sylvén, 2015; Rød & Calafato, 2025). In contrast, if learners' contact with the L2 and prior knowledge is limited, as was the case for L2 French, it may be more valuable to focus on expanding knowledge of high-frequent words which are most useful for the learners (He & Godfroid, 2019; Laufer & Nation, 2012).

Moreover, the substantial differences in productive vocabulary knowledge in L2 English and L2 French raise questions about the attainment targets in Flanders. Currently, the attainment targets for

English and French throughout secondary school are identical and even grouped under the same headings (as opposed to German, an L2 listed separately). By grade 8 of secondary school (12–13 years old), Flemish learners are expected to be at an A2 level according the Common European Framework of Reference (Council of Europe, 2020), in both French (four years of instruction) and English (one or two years of instruction). By the end of secondary school, in grade 12 (17–18 years old), Flemish learners are expected to have attained a B1 level in English and French. This means that Flemish learners are assumed to attain the same level of proficiency in English in a smaller instructional timeframe (in both years and hours), which can be explained in terms of the out-of-school ubiquity of English in Flanders, as is the case in many countries in Northern Europe (European Commission, 2012; Schmitt, 2019; Sundqvist, 2024)[38]. English also holds a status as a 'prestigious lingua franca' (Sundqvist, 2024, p. 6) and as the language of pop culture (Werner & Tegge, 2021), which adds a cultural dimension to the English learning context (Sockett, 2025). In view of the omnipresence and unique status of English, this PhD thesis seems to uncover a mismatch between the identical attainment targets for L2 English and L2 French and their highly disparate learning conditions and outcomes. At the level of education policy, it may be useful to explore whether differentiation of the attainment targets could better reflect the reality of the learning contexts.

7.4. Methodological contribution

The present PhD thesis makes a methodological contribution to L2 vocabulary and corpus research by combining several innovative features in terms of the research design, the data collection instruments and the analytical procedures. More specifically, the methodological contributions are situated in (a) the large sample of secondary school students and the cross-linguistic within-participants design; (b) the use of multiple writing tasks with a focus on informal writing; (c) the multifaceted analysis of vocabulary use; (d) the contributions in terms of pre-processing for the measurement of vocabulary use; (e) the adapted form recall vocabulary test with a strict scoring system; and (f) the measurement of out-of-school L2 contact through precise time estimates.

7.4.1. Sample and research design

The first methodological contribution relates to the investigated sample. This PhD project presented a large-scale investigation of L2

[38] It is useful to note that this ubiquity of English does not apply to all European countries or to other regions. For example, countries where TV is generally dubbed (e.g., France, Spain) are known to provide less English-rich environments than countries where TV tends to be subtitled (e.g., Sweden, Denmark) (Muñoz & Cadierno, 2021; Schurz & Sundqvist, 2022).

learners' vocabulary knowledge and use. Data were collected from a large sample with an initial total of 441 participants, compared to much smaller sample sizes in previous similar cross-sectional studies (e.g., N = 48 in Laufer, 1998; N = 56 in Henriksen & Danelund, 2015). In addition, while L2 research to date has mainly drawn from samples in university settings (e.g., Andringa & Godfroid, 2020; Plonsky, 2023), this project contributed to L2 research by focusing on secondary school students. Finally, great care was taken to ensure representativeness and generalizability of the findings. Although the sample was a convenience sample, as in most L2 research (DeKeyser et al., 2010; Plonsky, 2023), we paid careful attention to generalizability by including L2 learners in two grades, in both urban and rural areas, and in a variety of study tracks.

Furthermore, another asset of this PhD project was the cross-linguistic within-subject design, in which L2 learners each completed a battery of tests in both L2 English and L2 French. This is unlike previous studies, which have often investigated only one L2, usually English (e.g., Guo et al., 2021; Kusyk et al., 2025; Plonsky, 2023; Sundqvist, 2024). The inclusion of two L2s was important to enhance generalizability of the findings, as findings in one L2 cannot be automatically assumed to extend to other L2s (Treffers-Daller et al., 2008). Following a recent trend towards increased research on other L2s than English (e.g., Consolini & Kyle, 2024; Kisselev et al., 2025; Peters, Noreillie, et al., 2019; Sung et al., 2024; Treffers-Daller, 2013), the cross-linguistic design of this thesis was timely and important.

7.4.2. Writing tasks

Another methodological strength of this PhD thesis lies in the writing tasks. Specifically, the writing tasks were aimed at eliciting informal writing, a genre that has received limited attention in extant research (Heng et al., 2023; Weigle, 2010; Yu et al., 2023). There seems to be a tendency towards a focus on essay writing and in particular argumentative writing, both in research contexts as well as in pedagogical contexts, such as standardized writing assessment and writing education (Barkaoui & Knouzi, 2012; Heng et al., 2023; Weigle, 2010; Yoon & Tabari, 2023; Yu et al., 2023). As Olinghouse et al. (2012) state, learners' writing performance can depend on the writing genre, and without explicit attention to informal writing, learners' performance in this area may remain underexplored.

In addition, we administered two writing tasks in both L2 English and L2 French, following writing research emphasizing that no conclusions can be drawn about learners' writing ability based on a single writing sample (e.g., Bouwer et al., 2023; Olinghouse et al., 2012; Schoonen, 2005,

2012)[39]. Yet, research on vocabulary use in L2 writing has only rarely investigated multiple writing tasks completed by the same learners, and has most often drawn from learners' written output on a single task (Zenker & Kyle, 2021). This can be problematic, given that a few studies have indicated an effect of the writing prompt on learners' use of content words (Reid, 1990), use of varied words (Sung et al., 2025; Vandommele et al., 2017; Zenker & Kyle, 2021), and use of sophisticated words (Yoon, 2017). Consequently, this PhD thesis reduced the impact of the topic on lexical findings by administering two writing tasks in each L2.

The writing tasks and instructions were carefully designed to maximize the odds that learners' writing output would reflect their underlying competence. This is related to a common distinction in L2 research between *competence* and *performance*: performance is what is measured on L2 tests, and this is interpreted as a proxy for learners' underlying competence (Ellis, 2004). This can be linked to Cronbach's (1970) distinction between *ability* and *typical behavior*, as he posits that constructed-response tests (e.g., writing tasks) tend to elicit typical behavior that 'may not reflect [learners'] highest possible level of performance' (Schoonen, 2011, p. 707). Given these two distinctions, we took the following measures to elicit learners' true abilities. First, we carefully piloted several writing prompts prior to the data collection in order to ensure that the prompts were suitable for our sample (see also Chapter 3). We aimed for topics that would be of interest to the learners to enhance their motivation (Van Steendam et al., 2025), and we chose topics that were not conceptually difficult to write about so participants would be able to produce a sufficiently long text without losing interest or motivation. Second, although we did not inform the participants about the focus on vocabulary use in spite of a potential *had-I-known* effect (Schoonen, 2011), we explicitly urged the participants to complete the writing tasks to the best of their abilities. We did not emphasize vocabulary use in the instructions in order to promote spontaneous and authentic writing, a strategy in line with most previous research on vocabulary use in writing. Finally, we tried to reduce any potential anxiety by pointing out that learners' performance on the writing tasks would not count towards their school grades. Taken together, these measures increased the likelihood of learners' performance reflecting their underlying competence (Ellis, 2004) and ability (Cronbach, 1970)[40].

[39] For true generalizability, writing research has recommended administering three to five tasks for genre-specific conclusions, and three tasks in each genre for cross-genre conclusions (Bouwer et al., 2023).

[40] As Crossley (2020) notes, it remains difficult to determine if lexical diversity and sophistication in writing actually reflect learners' underlying vocabulary knowledge, or rather personal stylistic choices. He illustrates this in terms of lexical diversity in learners' output,

7.4.3. Lexical diversity and lexical sophistication indices

Methodological innovation was also reflected in the multifaceted analysis of vocabulary use in the L2 writing tasks. Vocabulary use was analyzed through two subconstructs of lexical richness: lexical diversity (the use of varied words) and lexical sophistication (the use of sophisticated words, based on word frequency). The distinct results for lexical diversity and lexical sophistication in this PhD thesis indicate the value of adopting a multivariate approach to learners' vocabulary use in writing (e.g., Consolini & Kyle, 2024). Moreover, while many studies have used either mean-based or band-based indices of lexical sophistication (e.g., Henriksen & Danelund, 2015; Kyle et al., 2018; Laufer, 1998), this thesis adopted both a mean-based approach (average log frequency; the average frequency of the lemmas used) as well as a band-based approach (lexical frequency profile; the proportion of lemmas beyond 2K).

The findings indicated that the average log frequency (ALF) may be more fine-grained and more sensitive to small changes in lexical sophistication than the lexical frequency profile (LFP). In Study 1, while the ALF measure detected a small cross-sectional difference in English mean-based lexical sophistication, no significant cross-sectional differences in band-based sophistication were observed by the LFP measure. ALF also showed a stronger relationship with English productive vocabulary knowledge than LFP. In Study 2, ALF in English was related to length of instruction and out-of-school contact, whereas LFP was not significantly predicted by length of instruction and only to a small extent by out-of-school contact. These different findings clearly show that the measures chosen determine to some extent which results will and can be observed (Lowie & Verspoor, 2015).

Although our findings supported Crossley et al.'s (2013) conclusion that mean-based measures may better capture small differences in lexical sophistication, this does not mean the LFP is not useful. Since the LFP reflects a proportion of words, it is less affected by outliers than the ALF (e.g., one very low-frequent or high-frequent word in a text can substantially affect the average). In addition, the LFP may lend itself better to pedagogical applications (Crossley et al., 2013), because it identifies the proportion of words from different frequency bands. This facilitates the use of the results in light of pedagogical goals and learner feedback, for instance, to track learners' progress across tasks or to provide additional vocabulary lists with mid-frequency vocabulary. In conclusion, a combination of mean- and band-based indices can provide the most comprehensive picture of learners' lexical sophistication, allowing

which may be an indication of their vocabulary knowledge, but it may also reflect their purposeful choices (not) to repeat words in light of text cohesion and personal writing style.

detection of small differences between learner texts as well as valuable insights applicable to L2 pedagogy.

Furthermore, the results regarding lexical diversity supported the reliability of MATTR as a measure that can distinguish between learners with different levels of vocabulary knowledge (e.g., Díez-Ortega & Kyle, 2024; Engber, 1995; Henriksen & Danelund, 2015; Zenker & Kyle, 2021). Study 1 showed that MATTR distinguished well between grade 10 and grade 12 learners in both L2 English and French, even though the learners in grade 12 had only had two years of additional L2 contact. Lexical diversity also had a moderate relationship with productive vocabulary knowledge in Studies 1 and 3, which seems to indicate that MATTR as a measure of lexical diversity reflects learners' form recall knowledge fairly well. Additionally, the findings also indirectly supported the reliability of MATTR across different text lengths (e.g., Bestgen, 2024; Kyle et al., 2024; Sung et al., 2024; Zenker & Kyle, 2021), given that learners wrote texts that varied greatly in the number of tokens (for English, the interquartile range was 387–542 tokens; for French, it was 195–391 tokens).

7.4.4. Pre-processing of learner texts

The present PhD thesis advanced methodological practices in corpus research by providing valuable insights into the impact of methodological decisions on the assessment of learners' vocabulary use in L2 English and L2 French. In particular, Study 3 demonstrated that consistency is key when comparing learner groups or L2s. Different choices regarding lemmatization (default or custom) and word inclusion (all words or content words) led to significantly different lexical findings in terms of lexical diversity, mean-based and band-based lexical sophistication. Although there was no one operationalization clearly related most strongly to productive vocabulary knowledge, the significantly different findings across operationalizations underscore the importance of methodological rigor, especially when comparing learner groups and L2s in the same study.

The findings especially highlighted the effect of the decision to include all words or content words only, thereby including or excluding function words from lexical analyses[41]. In our sample, function words were found to provide valuable information for lexical analyses of diversity and sophistication, indicating the value of including highly frequent and regularly repeated function words in the analyses. Study 3 showed a stronger relationship between productive vocabulary knowledge and

[41] Given that function words are highly frequent and they are necessary to express almost any (grammatically correct) sentence (Nation, 2013), it was to be expected that indices of lexical diversity and sophistication would be strongly affected by the exclusion of function words.

operationalizations based on all words than those based on content words, pointing to a slight advantage of indices based on all words (although the results sometimes varied depending on whether R^2 or AIC was considered). These results diverge somewhat from findings in previous studies, which have found that content word indices generally showed a slightly stronger relationship with L2 speaking and writing scores compared to all words, and content word indices were more consistently retained as meaningful predictors in regression models (Kyle & Crossley, 2015; Monteiro et al., 2020). The fact that in Study 3, the indices based on all words were better predicted by learners' vocabulary knowledge than those based on content words, may be linked to learners' limited productive vocabulary knowledge. It seems likely that the differences in diversity and sophistication between participants were less evident from their use of content words than their use of function words (see e.g., Zhang, 2022), which are virtually all among the 2,000 most frequent word families (Nation, 2013; Öztürk, 2022; Vilkaitė-Lozdienė & Schmitt, 2020).

Furthermore, the findings of Study 3 advanced our understanding of the effect of lemmatization on the measurement of lexical diversity and lexical sophistication. The findings indicated that the choice of the lemmatization scheme also significantly affected lexical findings. Default lemmatization seemed to be most strongly related to productive vocabulary knowledge and may thus better capture learners' lexical proficiency. This was against expectations, since the custom lemmatization scheme was primarily designed to account for the morphologically complex nature of French (Consolini & Kyle, 2024; Treffers-Daller, 2013; Treffers-Daller & Korybski, 2015). With our sample, a plausible explanation could be the low French proficiency of the participants, which may not have allowed the learners to exploit the lower-level categorization of custom lemmas based on mood and tense. For instance, when writing about what they would do if they won the lottery, learners might want to alternate between verb forms in the conditional tense, the future tense and the present tense. Under default lemmatization, all of these conjugated word forms would be classified as the same type, leading to more repetition and the same frequency count/band for each occurrence. In contrast, under custom lemmatization, each of the tenses would be classified under a different type, which could lead to higher diversity and higher sophistication. However, if the learners are not proficient enough to use these different tenses, they may opt for the present tense throughout their text, causing the custom lemmatization scheme to lose its benefit of accounting for varied word forms of different frequencies within one default lemma. It could therefore be hypothesized that higher-proficiency learners, who have more extensive knowledge of vocabulary (including inflected forms), may be able to better exploit the wide range of word forms, and they might show a stronger relationship between custom

lemmatization and productive vocabulary knowledge. This hypothesis is supported by the finding that custom lemmatization showed a stronger relationship with productive vocabulary knowledge in English than in French. Since English is the L2 for which the participants had more vocabulary knowledge (see Study 1), the custom lemmatization scheme may reflect learners' productive vocabulary knowledge better in English than French.

Additionally, unlike some previous studies (e.g., Laufer, 1998; Laufer & Nation, 1995), we adopted a strict approach to the lexical analyses wherein errors in the learner texts were not corrected, allowing us to capture learners' actual performance in written vocabulary use. This was in line with the strict scoring system for the form recall test. While the use of incorrectly spelled words may indicate partial knowledge or some level of familiarity (Laufer & Nation, 1995), it also reveals that learners are not fully able to use the words accurately in writing. Retaining the original words as they had been produced by the learners may therefore contribute to a more valid interpretation of learners' vocabulary use. Another strength of a strict scoring system is the relatively objective nature of this approach. A strict approach is straightforward in that no adjustments are made to the learner texts, thereby optimizing transparency and replicability. In contrast, a lenient system that accounts for partial knowledge by correcting spelling errors, requires subjective judgements. Researchers have to decide what is considered an error that should be corrected because it reflects sufficient knowledge by the learner (e.g., 'spelling errors that do not distort the word'; Laufer, 1998, p. 261), and what is considered a larger error indicating that learners are not able to produce or use the item independently[42].

A final methodological innovation in terms of pre-processing was the development of a natural language processing (NLP) pipeline in Python, which ensured consistency in the analyses across L2 English and L2 French. For English, several tools to analyze lexical diversity and sophistication were already available (Anthony, 2024; Cobb, n.d.; Kyle et al., 2018, 2021; McNamara et al., 2014; Meara & Bell, 2001)[43]. However, many of these tools are not open-source, which (a) limits transparency about the underlying analytical procedures, (b) complicates replicability of the findings, and (c) offers restricted control over methodological procedures, for instance, with regard to the lemmatization and the frequency lists. Moreover, many of the tools available for English only allow either a mean-based or a band-based approach to the analysis of lexical sophistication, which can compromise consistency across measures due to differences in

[42] When researchers opt for a lenient scoring system, it is important to report inter-rater reliability to account for this inherent subjectivity.

[43] These tools are (in order of the references): AntWordProfiler, Compleat Web VocabProfile, TAALES, TAALED, Coh-Metrix, and P_Lex.

the underlying analytical procedures of the tools. For French, on the other hand, limited tools were (and are) available to analyze written vocabulary use. In spite of recent efforts to create lexical analysis tools for French (e.g., Compleat Web VocabProfile, Cobb, n.d.; MultilingProfiler, Finlayson et al., 2023), their accessible web-based nature makes them more appropriate for pedagogical purposes. In addition, MultilingProfiler does not allow analyses of English texts, which was important given the cross-linguistic design of this PhD project.

As a result of the limited availability of tools allowing a multivariate approach to the analysis of lexical diversity and lexical sophistication in both L2 English and L2 French[44], we developed our own NLP pipeline based on Python packages by Kyle and colleagues (Kyle, 2025; Kyle et al., 2021). This allowed us to (a) have control over the exact analytical procedures and methodological decisions in the analyses, (b) remain consistent in the analyses across L2 English and French, and (c) share the Python code publicly for transparency.

7.4.5. Productive vocabulary test

Another methodological contribution of this PhD thesis was the use of a strict scoring system to score the productive vocabulary tests, in line with the strict approach to the analysis of vocabulary use. By defining productive vocabulary knowledge as full knowledge of the orthographic form that can be recalled based on a provided meaning(ful context), we ensured that the estimate of learners' productive vocabulary knowledge reflected only words fully known at the form recall level. Moreover, the strict approach benefited scoring reliability, since the intended word form was either produced fully correctly (and awarded a point) or not fully correctly (and not awarded a point). In this way, we avoided any subjective judgements of which orthographic errors were deemed to imply sufficient partial knowledge for the response to receive a point. For instance, the English target item *vulnerable* elicited a wide range of spelling errors, such as *vulnorable, volnurable, vunuarable, vounerbol, vonerable, vonerebol, volnourable, vounerable, vonrenabel, vonrounle, vownerable, vulnarebull*. With a strict scoring system, these responses are simply all scored as incorrect. With a lenient scoring system, however, raters have to judge which of the errors are deemed acceptable and which are not, often based on rather arbitrary criteria (e.g., a certain phonological resemblance, or a set limit of the number of incorrect letters).

[44] Recently, a new tool was published allowing flemma-based analyses of vocabulary use in English and French (Van Parys et al., 2025). While the tool is based on an NLP pipeline in Python, it is currently still only available through a demo version in an online web interface.

Additionally, unlike many studies that measure vocabulary knowledge at the level of meaning recognition, we used and adapted a form recall vocabulary test. There are several methodological benefits to the test format of the productive VocabLab tests (Noreillie et al., 2020; Peters, Velghe, & Van Rompaey, 2019), which measure knowledge of target items at the level of form recall. First, form recall tests require learners to demonstrate their knowledge by producing a word form. This counters a common critique of the checklist format (form recognition) where learners have to self-assess their knowledge (Read, 2020). Secondly, the recall-based format also reduces the risk of guessing commonly associated with recognition tests, as test takers cannot simply select a given response option (Nation, 2013; Read, 2020). Third, form recall tests measure knowledge of pre-determined target items sampled from frequency lists, and as such, they measure productive knowledge in a more controlled manner than free production tests (e.g., a word association task or a writing task).

For the present PhD thesis, some adjustments were made to improve the original VocabLab tests (see also Chapter 3). The first adaptation related to the L1 translations for each of the target items, which were already included in the French VocabLab test (Noreillie, 2019), but not in the English test (Jamers, 2021). An L1 translation was added at the end of each sentence of the English test, in line with Nation's (2013) recommendation to use bilingual tests if possible, and with a view to consistency in the format across English and French. The L1 translations were carefully selected based on translations in the online Van Dale dictionary and feedback by fellow researchers. A second adaptation concerned the elicited word form of the target items in the sentences, which were adjusted so the sentences would all elicit the base form (e.g., singular noun, infinitive verb). In this way, we avoided testing learners' morphological knowledge, which was not the focus of the vocabulary test. Finally, the number of items on the English test was reduced from 90 items to 60 items for consistency with the French test (30 items from 1–2K, 30 items from 3K). Based on data from the pilot study in this PhD thesis and a previous validation study (Jamers, 2021), 30 items were eliminated based on an item analysis with three criteria (item facility, discrimination index, and Cronbach's alpha if item deleted).

7.4.6. Questionnaire: out-of-school contact estimates

A final methodological contribution concerns the use of precise time estimates to measure out-of-school L2 contact. While most studies have used frequency-based questions with Likert scales (Sundqvist & Uztosun, 2023), we measured out-of-school contact through self-reported

estimates of the average time learners spent on out-of-school activities in hours and minutes per week (see e.g., Kaatari et al., 2023; Lu & Dang, 2023; Sundqvist & Wikström, 2015). These estimates allowed us to gain more detailed insight into learners' out-of-school contact than would have been the case with Likert scales. They also counter an important limitation of Likert scales, namely, how participants interpret relatively vague answer options like 'always', 'sometimes', or 'several times per week' (Sullivan & Artino, 2013; Sundqvist & Uztosun, 2023). The advantage of a weekly average over a daily average is that responses are less likely to be biased towards specific days of the week. For instance, learners' time investment might differ between weekdays versus weekends, or between days where they have other hobbies planned versus no other activities. As a result of this approach to measuring out-of-school contact, the time estimates provided a valuable approximation of learners' time investment in out-of-school activities, allowing us to capture variation between learners and to establish general patterns and trends.

7.5. Limitations and future directions

Although the empirical studies in this PhD thesis make a theoretical, pedagogical and methodological contribution to the field of SLA, they are not without limitations. The following sections address some of the limitations of the three studies with regard to the data collection instruments, the scoring procedure and lexical analyses, and the generalizability of the findings.

7.5.1. Questionnaire

One limitation is related to the questionnaire section used in Study 2 to gauge learners' out-of-school L2 contact. We asked learners to estimate the average time per week they spent on 11 out-of-school L2 activities. While this provided us with an informative approximation of learners' current average time investment, we did not gain any insight into their cumulative out-of-school contact over the years. Longitudinal studies could shed more light on the changes in out-of-school contact by tracking the same learners' out-of-school activities over time (see e.g., De Wilde, 2025; De Wilde & Lowie, 2024). We also did not investigate forms of engagement other than behavioral engagement (frequency/duration of the out-of-school contact). Researchers have called for the consideration of other aspects of engagement, like affective engagement (enjoyment during the L2 contact), cognitive engagement (attention paid to the activity), and linguistic engagement (focus on linguistic features) (e.g., Arndt, 2023; Fredricks et al., 2016; Sundqvist & Uztosun, 2023). This could be included in future studies in various ways through the use of questionnaires,

language diaries, event-based experience sampling, or interviews. Finally, while questionnaires are the most commonly used method of investigating out-of-school contact (Kusyk et al., 2025; Lee, 2022), some researchers have opted for different methods, like language diaries (e.g., Sylvén & Sundqvist, 2012) or experience sampling (e.g., Arndt et al., 2023). These methods may provide more detailed insight into learners' daily contact with the L2, although it has appeared to be challenging to mobilize learners to complete daily language diaries consistently (De Wilde, 2024; Sylvén & Sundqvist, 2012).

In terms of formal instruction, we only used the questionnaire data referring to learners' length of formal instruction in years. We have not analyzed the responses to questions about activities and input in the L2 classroom, such as the proportion of time that the teacher speaks in the L2, the number of writing tasks the learners had to complete, and to what extent different aspects of L2 learning (e.g., vocabulary, grammar, listening, writing, etc.) are addressed during instruction. Other valuable additions to this type of self-reported questionnaire data could include corpus analysis of class materials and textbooks (e.g., Bergström et al., 2025; Sun & Dang, 2020; Van Parys et al., 2024), or interviews with teachers (e.g., Bergström et al., 2022). Future research could also perform a latent profile analysis using data on formal instruction and out-of-school contact in order to create 'learner profiles' (e.g., high out-of-school contact and high in-class interaction), and investigate their relationship with learners' vocabulary knowledge and use (for an example of latent profile analysis, see e.g., Guo et al., 2025).

7.5.2. Writing tasks

The writing prompts in this PhD thesis required learners to write informal texts to a friend. Informal writing was chosen because this genre has received limited attention in previous research (Weigle, 2010; Yu et al., 2023), and the pilot study indicated that informal prompts generated more output and elicited more motivation from the learners than formal prompts. Although we believe the informal prompts allowed for diverse and sophisticated vocabulary use (see the L1 data in Study 1), it should be acknowledged that results may differ for other genres or registers (Bottini & Le Foll, 2025; Durrant et al., 2022; Durrant & Durrant, 2022; Vandommele et al., 2017). Since a subsample of the participants also completed two formal writing tasks in English (in addition to the informal writing tasks), it would be interesting to see whether the formal prompts may have elicited different levels of lexical diversity and sophistication from the same learners.

Given that the topic of writing prompts can in itself affect vocabulary use (e.g., Yoon, 2017; Zenker & Kyle, 2021), we administered two writing tasks in each L2 to counter these topic effects. However, for the lexical analyses, we treated the learners' writing output for the two topics as one text, because (a) the learners' writing output was sometimes limited (i.e., few tokens) and we aimed to discard as little of the data as possible, (b) we tried to keep the statistical models as parsimonious as possible, and (c) the inclusion of writing output based on two different topics was deemed to compensate to some extent for potential topic effects. Future studies could investigate the impact of topic effects on learners' vocabulary use in more detail. For instance, a recent small-scale error analysis of the data for this PhD project indicated that for French, learners tended to write longer texts for the 'complaint' prompt than for the 'prize' prompt (Stryckmans, 2025). It is possible that the topics might have also elicited different levels of lexical diversity and sophistication.

7.5.3. Scoring and lexical analyses

This PhD thesis adopted a strict interpretation of vocabulary knowledge and use in that responses on the form recall test were only awarded a point when they were spelled correctly, and spelling errors in the writing tasks were not corrected prior to the lexical analyses. This strict scoring system allowed us to capture learners' vocabulary knowledge of fully known words and their actual performance in written vocabulary use, and it ensured objectivity in the scoring procedure. Nevertheless, it is possible that the findings of this PhD thesis may have been affected by the strict scoring system. A lenient, more sensitive, scoring system can reward partial knowledge of the learners by awarding a point to misspelled vocabulary test responses and by correcting spelling errors in writing output. This could provide a different perspective on learners' productive vocabulary knowledge and vocabulary use. However, a lenient scoring system also comes with its own challenges, such as the inherent subjectivity in deciding which errors deserve a point on a vocabulary test and which errors deserve to be corrected in writing. These are not straightforward decisions, and it is important for researchers to report inter-rater reliability when they do adopt a lenient scoring system.

Further, the measures of lexical sophistication used in the three studies were frequency-based. While frequency is a widely used criterion to determine how difficult or advanced a word is (e.g., Ellis, 2002; Kyle, 2020b) and a strong predictor of L2 vocabulary knowledge and L2 proficiency (e.g., Boone & De Wilde, 2023; Kyle et al., 2018; Nation, 2006), researchers are increasingly recognizing the multi-dimensional nature of lexical sophistication. Interpretations of lexical sophistication have been

expanded with measures related to range, contextual diversity, *n*-gram properties (e.g., bigram strength of association) and psycholinguistic norms (e.g., concreteness, familiarity, age of acquisition) (e.g., Hashimoto & Egbert, 2019; Kim et al., 2018; Kyle et al., 2018; Nicklin et al., 2025). In future research, it would be useful to continue to adopt a multi-dimensional approach to lexical sophistication, beyond the traditional frequency measures. While this is increasingly common in research on L2 English, research on other L2s seems to be lagging behind, partly because of a lack of available tools to analyze other L2s. While no tools currently seem to exist that allow a multi-dimensional perspective to lexical sophistication in French, tools like Compleat Web VocabProfile (Cobb, n.d.), MultilingProfiler (Finlayson et al., 2023) and LexPro (Van Parys et al., 2025) can provide teachers and researchers with useful tools to analyze word frequency in French in small-scaled analyses[45].

7.5.4. Generalizability

Finally, this PhD thesis provided insights into productive vocabulary knowledge and vocabulary use of learners of L2 English and L2 French in Flanders, where Dutch is the official language and medium of instruction. We aimed to increase generalizability of the findings by investigating two L2s in a within-participants design and in the same geographical region. However, it remains to be seen to what extent these findings can be generalized to other geographical or learning contexts. For instance, we measured out-of-school contact in the context of Flanders, a subtitling region. The results on out-of-school contact may differ depending on certain cultural habits such as subtitling or dubbing (Muñoz & Cadierno, 2021). Similarly, the findings may not necessarily generalize beyond the European context, or to other typologically different languages, especially if the L1 and L2 have a different writing system (e.g., Japanese). Nonetheless, the differential findings in this PhD thesis for L2 English and French within the same learners, highlight the importance of not assuming that research findings can automatically be generalized from one L2 to another (Treffers-Daller et al., 2008). It is important to continue research on learners' productive vocabulary knowledge and vocabulary use in various geographical contexts, learning settings, and in other L2s in addition to English.

7.6. Conclusion

This PhD thesis constitutes the first large-scale study of secondary school learners' productive L2 vocabulary knowledge and vocabulary use

[45] Note that batch processing is currently not possible in any of these tools.

in L2 writing in two target languages, L2 English and L2 French. Through three empirical studies, the thesis presented a multi-dimensional investigation of L2 learners' form recall knowledge, lexical diversity, mean-based and band-based lexical sophistication, while also examining the impact of learner-related variables (grade, out-of-school L2 contact, length of formal instruction) and methodological factors (lemmatization, word inclusion).

This PhD project has theoretical implications for SLA and L2 vocabulary research, pedagogical implications for L2 teachers, materials developers and policy-makers, and methodological implications for L2 vocabulary and corpus research. The results demonstrated that (a) productive vocabulary knowledge, lexical diversity and lexical sophistication may develop at different rates, with lexical sophistication developing more slowly; (b) the development of productive vocabulary knowledge, lexical diversity and lexical sophistication is driven by both formal instruction and out-of-school L2 contact, which underscores the importance of L2 contact; (c) methodological choices regarding lexical indices, lemmatization and word inclusion have a significant impact on lexical findings; and (d) results can vary considerably depending on the L2 being measured, even within the same learners. The findings highlight the value of adopting a multi-dimensional approach by examining multiple aspects of vocabulary knowledge and use, employing several indices of lexical diversity and sophistication, and investigating multiple L2s.

This thesis provides a foundation for further research aimed at enhancing our understanding of L2 learners' productive vocabulary knowledge and vocabulary use. While the thesis contributes to a growing body of L2 research, much remains to be explored in terms of vocabulary development, and in particular vocabulary use in writing. There is a need for more research adopting multivariate approaches to measuring vocabulary use, and future research may also want to focus on how to enhance learners' lexical diversity and sophistication in writing (see e.g., Schmitt, 2020). Finally, given the differences between L2 English and L2 French demonstrated in this thesis, it is essential that future research continues to explore not only L2 learning in English, but also in other languages. Ultimately, this thesis adds some pieces to the puzzle, but there is a long way to go towards full understanding of how productive vocabulary knowledge and vocabulary use develop. In other words, there is still much knowledge waiting to be uncovered, or perhaps… *revealed? unearthed? excavated?* Well, I guess like the learners who took part in this PhD project, I too find myself asking: *what's the word*?

Final remarks

This section marks not only the end of a written account of this PhD project, but also the end of a real-life four-year trajectory. And what a trajectory that was. I am eternally grateful that I was given this opportunity to spend **four years learning, discovering, and growing**. It truly was a privilege. I would like to use this section to reflect on everything I have done over the course of the last four years. I recently read a quote online that said: 'Research is spending 6 hours reading 35 papers, so you can write 1 sentence containing 2 references' (Bryan Gaensler). And while that is absolutely true (☺), luckily I was also able to experience that being a PhD student is so much more than that.

During the past four years, **I have worn many hats**. I have been a researcher, a student, a colleague, a writer, a revisor, a presenter, a reviewer, a proofreader, a traveler, a data coder, an analyst, a lecturer, a supervisor, an evaluator, a staff representative, a member of working groups and task forces, an event organizer, and probably even more that I currently cannot recall (a recognition format might have made this list easier to compose). All in all, my various roles can be grouped under three main categories: research, teaching and service.

The research hat

I spent these four years mainly as a researcher. For my PhD, I set up and conducted a **large-scale data collection** with L2 learners in two grades of secondary school who completed a test battery in both L2 English and L2 French. This led to a large **cross-sectional and cross-linguistic learner corpus**, containing writing output in two L2s from over 400 participants (N = 1692 texts). This learner corpus is accompanied by extensive metadata, such as the participants' grade, gender, vocabulary test scores, and rich questionnaire data.

I have been able to disseminate my research findings through **publications in high-quality international peer-reviewed journals**. My first PhD study was published in the top-tier journal *Applied Linguistics* (five-year impact factor 4.8). My two other PhD studies are currently under review at high-quality journals with a high impact factor. I also value dissemination to non-academic audiences through **science communication**, which is evidenced in the fact that I have published accessible OASIS summaries of my research and I co-organized a workshop at Kinderuniversiteit (2024).

In addition to my PhD studies, I have also been involved in **two other publications**. One of these is a collaboration with Prof. Eva Puimège

and Prof. Elke Peters about pretesting effects on incidental vocabulary learning. This study, which allowed me to expand my methodological skills to experimental methods, was recently published in *Studies in Second Language Acquisition*[46] (five-year impact factor 5.5). I am also the first author of a publication on primary school children's out-of-school language contact and vocabulary knowledge prior to formal instruction, a collaboration with Lieven Bollansée, Prof. Eva Puimège, and Prof. Elke Peters. This study has been submitted for an upcoming special issue in *ITL – International Journal of Applied Linguistics* (impact factor 2.3). It has been a true pleasure to collaborate with fellow members of the Language, Education and Society research group at our university.

Furthermore, I have been actively building a solid (international) research network. I have presented my research at several of the most **prominent national and international conferences** (e.g., AAAL, EuroSLA, BAAL Vocab SIG, Earli SIG Writing, BAAHE). I also attach great importance to international mobility, which is evidenced by my **research stay at the University of Oregon**. I was invited to give a talk at the Learner Corpus Research and Applied Data Science lab by Prof. Kristopher Kyle, a leading authority with regard to the analysis of lexical and lexico-grammatical features of written and spoken L2 texts. I **secured a travel grant** from the Research Foundation – Flanders (FWO) to fund this study visit abroad. I am truly grateful for all of the opportunities these four years have given me to explore the world and meet so many interesting people.

Moreover, during my time as a PhD *researcher*, I was also a PhD *student*. I spent the past four years enhancing my research expertise and acquiring **theoretical and methodological skills**. I have not only gained knowledge of research into SLA, L2 vocabulary and L2 writing, I have also expanded my horizons by acquiring skills in **corpus analysis** and **statistical methods**. I gained experience with corpus tools like TAALES, TAALED, MultilingProfiler, Compleat Web VocabProfile, and I also developed a natural language processing pipeline in Python to analyze the learner corpus I collected. For the data collections, I acquainted myself with tools like Gorilla Experiment Builder, Prolific and Clickworker. In addition, I familiarized myself with statistical methods like mixed-effects modeling, structural equation modeling, and generalized additive modeling, and I was able to work in both JASP and R to bring my statistical analyses to a satisfactory conclusion.

[46] Puimège, E., Caltabellotta, E., & Peters, E. (2025). Pretesting effects on incidental L2 vocabulary learning through reading. *Studies in Second Language Acquisition*, *47*(3), 774–793. https://doi.org/10.1017/S0272263125100910

The teaching hat

In addition to research, I have taken up a range of teaching duties over the past four years. First, I have taught **guest lectures** on writing assessment (*Master of Advanced Studies in Linguistics*) as well as on academic writing, pragmatics, translation and interpreting (*Bachelor of Applied Language Studies*). In 2023–2024, I **co-taught the course** 'Methods for Applied Language Studies' (*Bachelor of Applied Language Studies*), during which I taught classes on various topics related to research methodology. I have always loved teaching (little me made my mother take my classes, to give you an idea), and I am very grateful that I was trusted with this responsibility during my PhD.

Further, I have **(co-)supervised two bachelor's theses and three master's theses** throughout my PhD. This allowed me to guide students in their academic research and writing process on a day-to-day basis, and to witness their development as researchers. I greatly enjoyed watching and helping them grow and expand their research skills.

Moreover, I have been able to support numerous colleagues by **developing and correcting exams**. Of course, I developed and corrected my own exam questions for the guest lectures and course I taught. In addition, since the start of my PhD, I have collaborated yearly with the colleagues teaching 'English proficiency 2' in the Bachelor program by developing two vocabulary tests. In 2021–2022, I also provided students with feedback during in-class writing assignments and scored the online exercises they completed. In 2024–2025, I contributed to the program *Legal translation and interpreting* by correcting the written entry exams for the French language strand.

The service hat

Apart from research and teaching, I happily spent three years of my PhD contributing to university policy matters. For the past three years, I have been (and currently still am) an **ABAP representative**[47], which means I represent the junior academic staff. I have represented the ABAP at three different policy levels: (a) university-level, in the university-wide ABAP meetings, (b) faculty-level, in the Faculty of Arts Council and Faculty ABAP Council, and (c) campus-level, in the Antwerp Campus Council and the Campus Forum for Research. In 2023–2024, I was elected **vice-chair** of the ABAP Council.

[47] ABAP stands for 'assisterend en bijzonder academisch personeel' ('assisting and special academic staff'). This refers to staff with a temporary contract, such as PhD researchers, postdoctoral researchers, research associates, and teaching assistants.

Additionally, I have taken part in a faculty **task force** on transgressive behavior (2022) and I am a member of the task force for suicide prevention policy on the Antwerp campuses (2023–present). I have also been an active member of the **Faculty Diversity Network** since its creation in 2022. My position in all of these policy bodies has allowed me to think critically about existing and new policies, and to collaborate with colleagues at the university in order to pursue a healthy, safe and stimulating work environment.

Finally, alongside my involvement in policy matters, I have also contributed to the academic community in more informal ways. In 2023, I was a member of the **organizing committee** of the *Linguistics PhD Day* at KU Leuven. I also co-organized the yearly *How Do You Do (It)?* event aimed at the ABAP staff in 2024. In addition, I have been an **ad-hoc reviewer** for well-established journals in the field (*Language Learning, International Review of Applied Linguistics in Language Teaching, ITL – International Journal of Applied Linguistics, Journal of Child Language*) and the international *Second Language Research Forum* conference in 2025. Finally, I have strived to make new colleagues feel welcome and find their way around the complex entity of KU Leuven by serving as a **'buddy' for new PhD researchers** at our research unit and providing a warm welcome to new ABAP members on campus Antwerp.

A final note

Writing this section has made it abundantly clear to me (even more so than it already was) what an incredible experience this has been. The variety of roles and responsibilities I was able (and allowed) to take up, is truly unique. It is amazing that I was financially compensated to learn so many new things, to travel the world and to get to know other researchers with such interesting perspectives. I was able to do this under the guidance of three researchers with ample expertise in their field, each with their own perspective and point of view. Although, let it be known that it turned out to be quite the challenge to incorporate three people's (sometimes very different) feedback into my own writing ☺.

These four years have been challenging, but also incredibly rewarding. I have grown in ways that I never could have imagined. It is safe to say that I absolutely did not understand what I was getting into when I started this PhD journey, but I am grateful for everything it has brought me. I am excited to see where the future takes me.

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, *21*(3), 199–226. https://doi.org/10.1080/09588220802090246

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*(4), 721–754. https://doi.org/10.14746/ssllt.2018.8.4.2

Almeida, J., Knobel, M., Finkbeiner, M., & Caramazza, A. (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review*, *14*(6), 1177–1182. https://doi.org/10.3758/BF03193109

Alonso Alonso, R. (2022). Out-of-school contact with L2 English: Perspectives from language learners. *RAEL: Revista Electrónica de Lingüística Aplicada*, *21*(1), Article 1.

Amengual-Pizarro, M., & Alonso Alonso, R. (2024). *Informal language learning: Expanding the English learning horizons beyond the classroom*. Peter Lang.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Anthony, L. (2024). *AntWordProfiler* (Version v.2.2.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntWordProfiler

Anthony, L., & Nation, I. S. P. (2021). *Picture Vocabulary Size Test* (Version 1.2.3) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/pvst/releases/PVST123/help.pdf

Arndt, H. L. (2019). *Informal second language learning: The role of engagement, proficiency, attitudes, and motivation* [Doctoral dissertation]. University of Oxford.

Arndt, H. L. (2023). Construction and validation of a questionnaire to study engagement in informal second language learning. *Studies in Second Language Acquisition*, *45*(5), 1456–1480. https://doi.org/10.1017/S0272263122000572

Arndt, H. L., Granfeldt, J., & Gullberg, M. (2023). The Lang-Track-App: Open-source tools for implementing the experience sampling method in second language acquisition research. *Language Learning*, *73*(3), 869–903. https://doi.org/10.1111/lang.12555

Avello, D., & Muñoz, C. (2025). Vocabulary learning from captioned video viewing in primary school learners: The influence of word- and context-related factors. In A. Pattemore & F. Gesa (Eds.), *Foreign language learning from audiovisual input: The role of original version television* (pp. 17–37). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-91001-2_2

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new Computer Adaptive Test of Size and Strength (CATSS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

Bardovi-Harlig, K. (2020). One functional approach to L2 acquisition: The concept-oriented approach. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (3rd ed., pp. 40–63). Routledge.

Barkaoui, K., & Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in writing assessments. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 55–82). Brill. https://doi.org/10.1163/9789004248489_002

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Batista, R., & Horst, M. (2016). A new receptive vocabulary size test for French. *The Canadian Modern Language Review*, *72*(2), 211–233.

Baudot, J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain*. Les Presses de l'Université de Montréal.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Benson, P. (2013). *Teaching and researching autonomy* (2nd ed.). Routledge.

Benson, P., & Reinders, H. (2011). *Beyond the language classroom*. Springer.

Bergström, D. (2024). 'Solve the crossword': An analysis of task design in EFL materials from a vocabulary perspective. *The Language Learning Journal*, *52*(4), 426–440. https://doi.org/10.1080/09571736.2023.2193833

Bergström, D., Norberg, C., & Nordlund, M. (2022). "Words are picked up along the way" – Swedish EFL teachers' conceptualizations of vocabulary knowledge and learning. *Language Awareness*, *31*(4), 393–409. https://doi.org/10.1080/09658416.2021.1893326

Bergström, D., Norberg, C., & Nordlund, M. (2025). Do textbooks support incidental vocabulary learning? – A corpus-based study of Swedish intermediate EFL materials. *Education Inquiry*, *16*(1), 69–87. https://doi.org/10.1080/20004508.2022.2163050

Berthele, R., Lenz, P., & Peyer, E. (2022). Predicting foreign language skills based on first languages: The role of lexical distance and relative morphological complexity. *Poznan Studies in Contemporary Linguistics*, *58*(3), 419–448. https://doi.org/10.1515/psicl-2022-0020

Berthele, R., & Udry, I. (Eds.). (2021). *Individual differences in early instructed language learning: The role of language aptitude, cognition, and motivation*. Language Science Press. https://doi.org/10.5281/zenodo.5378471

Bestgen, Y. (2024). Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, *74*(3), 638–671. https://doi.org/10.1111/lang.12630

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9–37. https://doi.org/10.1515/cllt-2012-0002

Blanchard, M. (2021). *Pragmatic markers in native and non-native Englishes: A study into the use of and attitudes to pragmatic markers* [Doctoral dissertation]. KU Leuven.

Bollansée, L., Puimège, E., & Peters, E. (2021). "Watch out! Behind you is the enemy!" An exploratory study into the relationship between extramural English and productive vocabulary knowledge. In V. Werner & F. Tegge (Eds.), *Pop culture in language education* (pp. 199–214). Routledge.

Bonin, P., Fayol, M., & Gombert, J.-E. (1998). An experimental study of lexical access in the writing and naming of isolated words. *International Journal of Psychology*, *33*(4), 269–286. https://doi.org/10.1080/002075998400312

Boone, G., & De Wilde, V. (2023). Productive vocabulary knowledge in L2 German: Which word-related variables matter? *System*, *118*, 103150. https://doi.org/10.1016/j.system.2023.103150

Bottini, R., & Le Foll, E. (2025). The more proficient the learners, the less sophisticated their L2 vocabulary?: The curious effect of the reference corpus on mean-frequency measures of lexical sophistication. *International Journal of Learner Corpus Research*, *11*(1), 47–78. https://doi.org/10.1075/ijlcr.23029.bot

Bouwer, R., Van Steendam, E., & Lesterhuis, M. (2023). Guidelines for the validation of writing assessment in intervention studies. In F. De Smedt, R. Bouwer, T. Limpo, & S. Graham (Eds.), *Conceptualizing, designing, implementing, and evaluating interventions* (pp. 199–223). Brill.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 21–46). John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., Housen, A., & Pallotti, G. (2025). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, *75*(2), 533–574. https://doi.org/10.1111/lang.12669

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, *91*, 102246. https://doi.org/10.1016/j.system.2020.102246

Busby, N. L. (2021). Words from where?: Predictors of L2 English vocabulary among Norwegian university students. *ITL - International Journal of Applied Linguistics*, *172*(1), 58–84. https://doi.org/10.1075/itl.19018.bus

Bybee, J. (1998). The emergent lexicon. *Proceedings of the Panels of the Chicago Linguistic Society's Thirty-Fourth Meeting*, 421–435. https://ling.unm.edu/assets/documents/bybee1998emergentlexicon.pdf

Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., & Peters, E. (under review). The role of out-of-school language contact and instruction in L2 English and L2 French vocabulary knowledge and use.

Caltabellotta, E., Van Steendam, E., Noreillie, A.-S., Puimège, E., Creten, S., & Peters, E. (2025). Vocabulary knowledge and vocabulary use in writing: A cross-sectional comparison of L2 English and L2 French. *Applied Linguistics*, *Advance online publication*, amaf009. https://doi.org/10.1093/applin/amaf009

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *The Psychological Record*, *2*, 379–386.

Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.

Chen, X., Dong, Y., & Yu, X. (2018). On the predictive validity of various corpus-based frequency norms in L2 English lexical processing. *Behavior Research Methods*, *50*(1), 1–25. https://doi.org/10.3758/s13428-017-1001-8

Chenu, F., & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition et Interaction En Langue Étrangère*, *1*, 17–38. https://doi.org/10.4000/aile.4506

Clenton, J., & Booth, P. (Eds.). (2022). *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary*. Routledge. https://doi.org/10.4324/9780429285400

Cobb, T. (n.d.). *Compleat Web VocabProfile* (Version v.2.6) [Computer software]. https://www.lextutor.ca/vp/comp/

Cobb, T., & Horst, M. (2011). Does Word Coach coach words? *CALICO Journal*, *28*(3), 639–661.

Conklin, K. (2020). Processing single-word and multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 174–188). Routledge.

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press. https://doi.org/10.1017/9781108233279

Consolini, C. H., & Kyle, K. (2024). The relationship between L2 Spanish proficiency and features of written lexical and lexicogrammatical use. *Applied Linguistics*, *45*(6), 992–1009. https://doi.org/10.1093/applin/amae032

Council of Europe. (2020). *Common European Framework of Reference for Languages: Companion Volume*. Council of Europe Publishing.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Cribari-Neto, F., & Lima, M. da G. A. (2014). New heteroskedasticity-robust standard errors for the linear regression model. *Brazilian Journal of Probability and Statistics*, *28*(1), 83–95.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). Harper and Row.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), Article 3. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*(4), 965–981. https://doi.org/10.1016/j.system.2013.08.002

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/1f21q8ck

Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal*, *100*(3), 702–715. https://doi.org/10.1111/modl.12344

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/APPLIN/AMT056

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Cuetos, F., Alvarez, B., González-Nosti, M., Méot, A., & Bonin, P. (2006). Determinants of lexical access in speech production: Role of word frequency and age of acquisition. *Memory & Cognition*, *34*(5), 999–1010. https://doi.org/10.3758/BF03193247

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In M. H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234–244). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.016

Daller, M. H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In M. H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150–164). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.011

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). John Benjamins.

Davies, M. (2008). *Corpus of Contemporary American English* [Dataset]. http://corpus.byu.edu/coca/

Day, R., & Robb, T. (2015). Extensive reading. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 3–12). Routledge.

De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *EUROSLA Yearbook*, *15*(1), 69–94. https://doi.org/10.1075/eurosla.15.03dec

De Clercq, B., & Simoens, H. (2015). Two at the time: The parallel learning of French and English as foreign languages. *Círculo De Lingüística Aplicada a La Comunicación*, *63*, 9–41. https://doi.org/10.5209/rev_CLAC.2015.v63.50168

De Clercq, M. (2016). *Tester l'étendue du vocabulaire français des étudiants néerlandophones: Une étude comparative des versions monolingues et bilingues d'un test réceptif et productif* [Unpublished master's thesis]. KU Leuven.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1). https://doi.org/10.1017/S0272263111000489

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, *68*(4), 906–941. https://doi.org/10.1111/lang.12296

De Wilde, V. (2024). Research methods for investigating young learners' engagement with foreign languages outside of school. *Research Methods in Applied Linguistics*, *3*(3), 100159. https://doi.org/10.1016/j.rmal.2024.100159

De Wilde, V. (2025). Adolescent learners' L2 English vocabulary knowledge and contact with extramural English: Longitudinal development and relationships between L2 vocabulary and extramural English. *Bilingualism: Language and Cognition*, *First View*, 1–12. https://doi.org/10.1017/S1366728925100540

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020a). Learning English through out-of-school exposure: How do word-related variables and proficiency influence receptive vocabulary learning? *Language Learning*, *70*(2), 349–381. https://doi.org/10.1111/LANG.12380

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020b). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, *23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Young learners' L2 English after the onset of instruction: Longitudinal development of L2 proficiency and the role of individual differences.

Bilingualism: Language and Cognition, *24*(3), 439–453. https://doi.org/10.1017/S1366728920000747

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2022). Formal versus informal L2 learning. How do individual differences and word-related variables influence French and English L2 vocabulary learning in Dutch-speaking children? *Studies in Second Language Acquisition*, *44*, 87–111. https://doi.org/10.1017/S0272263121000097

De Wilde, V., & Lowie, W. (2024). The forest and the trees: Investigating groups and individuals in longitudinal second language English speaking development. *Language Learning*, *75*(4), 988–1018. https://doi.org/10.1111/lang.12698

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(3), 413–438. https://doi.org/10.1017/S0142716410000056

Dewaele, J.-M. (Ed.). (2005a). *Focus on French as a foreign language: Multidisciplinary approaches*. Multilingual Matters.

Dewaele, J.-M. (2005b). Sociodemographic, psychological and politicocultural correlates in Flemish students' attitudes towards French and English. *Journal of Multilingual and Multicultural Development*, *26*(2), 118–137. https://doi.org/10.1080/01434630508668400

Dewaele, J.-M., Botes, E., & Greiff, S. (2023). Sources and effects of foreign language enjoyment, anxiety, and boredom: A structural equation modeling approach. *Studies in Second Language Acquisition*, *45*(2), 461–479. https://doi.org/10.1017/S0272263122000328

Dewaele, J.-M., & Macintyre, P. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237–274. https://doi.org/10.14746/ssllt.2014.4.2.5

Dewaele, J.-M., & Saito, K. (2024). Are enjoyment, anxiety and attitudes/motivation different in English foreign language classes compared to LOTE classes? *Studies in Second Language Learning and Teaching*, *14*(1), Article 1. https://doi.org/10.14746/ssllt.42376

Dewaele, J.-M., Saito, K., & Halimi, F. (2023). How foreign language enjoyment acts as a buoy for sagging motivation: A longitudinal investigation. *Applied Linguistics*, *44*(1), 22–45. https://doi.org/10.1093/applin/amac033

Díez-Ortega, M., & Kyle, K. (2024). Measuring the development of lexical richness of L2 Spanish: A longitudinal learner corpus study. *Studies in Second Language Acquisition*, *46*(1), 169–199. https://doi.org/10.1017/S0272263123000384

Divjak, D., & Caldwell-Harris, C. L. (2015). Frequency and entrenchment. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 53–75). De Gruyter. https://doi.org/10.1515/9783110292022-004

Dörnyei, Z. (2005). Motivation and "self-motivation." In Z. Dörnyei, *The psychology of the language learner: Individual differences in second language acquisition* (pp. 65–119). Routledge.

Dörnyei, Z., & Csizér, K. (2012). How to design and analyze surveys in second language acquisition research. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 74–94). Wiley-Blackwell.

Dunn, L. M., Dunn, L. M., & Thériault-Whalen, C. M. (1993). *Échelle de vocabulaire en images Peabody: EVIP*. Psycan.

Durbahn, M., Rodgers, M. P. H., Macis, M., & Peters, E. (2024). Lexical coverage in L1 and L2 viewing comprehension. *Studies in Second Language Acquisition*, *46*(4), 1045–1068. https://doi.org/10.1017/S0272263124000391

Durbahn, M., Rodgers, M. P. H., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, 102166. https://doi.org/10.1016/j.system.2019.102166

Durrant, P., & Durrant, A. (2022). *Appropriateness* as an aspect of lexical richness: What do quantitative measures tell us about children's writing? *Assessing Writing*, *51*, 100596. https://doi.org/10.1016/j.asw.2021.100596

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in first-year composition assignments. *International Journal of Corpus Linguistics*, *24*(1), 33–66. https://doi.org/10.1075/ijcl.17052.dur

Durrant, P., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research methods in vocabulary studies*. John Benjamins. https://doi.org/10.1075/rmal.2

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, *16*(2), 227–252. https://doi.org/10.1177/1362168811431377

Educational Testing Service. (n.d.). *Score user guide for the TOEIC*. https://www.ets.org/pdfs/toeic/toeic-speaking-writing-score-user-guide.pdf

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272. https://doi.org/10.1177/0265532212459028

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language

acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function introduction to the special issue. *The Modern Language Journal*, *93*(3), 329–335. https://doi.org/10.1111/j.1540-4781.2009.00893.x

Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (3rd ed., pp. 63–82). Routledge.

Ellis, R. (2004). *Understanding second language acquisition*. Oxford University Press.

Ellis, R. (2017). Task-based language teaching. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 108–125). Routledge.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

European Commission. (2012). *Special Eurobarometer 386: Europeans and their languages*. https://d1wqtxts1xzle7.cloudfront.net/38342278/ebs_386_en-libre.pdf

Eurostat. (2024). *Foreign language learning statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Foreign_language_learning_statistics

Eyckmans, J. (2004). *Measuring receptive vocabulary size. Reliability and validity of the Yes/No vocabulary test for French-speaking learners of Dutch* [Doctoral dissertation]. Catholic University of Nijmegen.

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal*, *31*(2), 105–119. https://doi.org/10.1177/003368820003100205

Fan, M. (2009). An exploratory study of collocational use by ESL students – A task based approach. *System*, *37*(1), 110–123. https://doi.org/10.1016/j.system.2008.06.004

Finlayson, N., Marsden, E., & Anthony, L. (2023). Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, *118*, 103122. https://doi.org/10.1016/j.system.2023.103122

Fonteyne, M., Montero Perez, M., Daems, J., & Macken, L. (2025). Selecting instruments for investigating L2 Swedish writing data of L1 Dutch learners across proficiencies. *Nordand*, *20*(1), 1–17. https://doi.org/10.18261/nordand.20.1.1

Forsberg-Lundell, F., & Lindqvist, C. (2012). Vocabulary aspects of advanced L2 French: Do lexical formulaic sequences and lexical

richness develop at the same rate? *Language, Interaction and Acquisition*, *3*(1), 73–92. https://doi.org/10.1075/lia.3.1.05for

Forsman, L. (2004). *Language, culture, and context: Exploring knowledge and attitudes among Finland-Swedish EFL-students with particular focus on extracurricular influence* [Unpublished Master's thesis]. Åbo Akademi University.

Forster, K. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). North Holland.

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, *43*, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002

Fulcher, G. (2024). *Practical language testing* (2nd ed.). Routledge. https://doi.org/10.4324/9781003373629

Gablasova, D., & Brezina, V. (2021). Words that matter in L2 research and pedagogy: A corpus-linguistics perspective. *Studies in Second Language Acquisition*, *43*(5), 958–961. https://doi.org/10.1017/S027226312100070X

Ghamarian-Krenn, K. (2023). *Extramural English and academic verb knowledge: A longitudinal study of Viennese students majoring in English* [Doctoral dissertation]. University of Vienna. https://doi.org/10.25365/thesis.63632

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing: Expanding Nation's framework. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). Routledge.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, *21*(3), 563–584. https://doi.org/10.1017/S1366728917000219

González-Fernández, B. (2021). Vocabulary knowledge: Lexical depth and its relationship with out-of-class exposure. In P. Szudarski & S. Barclay (Eds.), *Vocabulary theory, patterning and teaching* (pp. 19–45). Multilingual Matters.

González-Fernández, B. (2025). How is vocabulary learnt? An acquisitional sequence of L2 word knowledge. *TESOL Quarterly*, *59*(2), 755–784. https://doi.org/10.1002/tesq.3342

González-Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *ITL - International Journal of Applied Linguistics*, *166*(1), 94–126. https://doi.org/10.1075/itl.166.1.03fer

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. Carnegie Corporation of New York.

Granfeldt, J., Erickson, G., Bardel, C., Sayehli, S., Ågren, M., & Österberg, R. (2023). Speaking French, German and Spanish in Swedish lower secondary school: A study on attained levels of proficiency. *Apples - Journal of Applied Language Studies*, *17*(2), Article 2. https://doi.org/10.47862/apples.127819

Green, R. (2013). *Statistical analyses for language testers*. Palgrave Macmillan.

Green, R. (2019). Item analysis in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume I* (pp. 15–29). Routledge.

Griffin, G. F. (1992). *Aspects of the psychology of second language vocabulary list learning* [Doctoral dissertation]. University of Warwick. http://webcat.warwick.ac.uk/record=b1449317~S1

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses universitaires de France.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Guo, Q., Zhou, X. (Emily), & Gao, X. (Andy). (2021). Research on learning and teaching of languages other than English in *System*. *System*, *100*, 102541. https://doi.org/10.1016/j.system.2021.102541

Guo, S., Yao, H., & Lee, J. S. (2025). A latent profile analysis of Informal Digital Learning of English: Emotion regulation strategies as predictors and willingness to communicate as outcomes. *System*, *132*, 103686. https://doi.org/10.1016/j.system.2025.103686

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558–579. https://doi.org/10.1177/0265532220979562

Gyllstad, H., Sundqvist, P., Peters, E., Rindal, U., Skar, G. B., & Ulfat, N. (2025). Young learners' receptive L2 English vocabulary knowledge in relation to extramural English exposure at the onset of formal instruction in Norway. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0289

Hair, J. F., & Alamer, A. (2022). Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. *Research Methods in Applied Linguistics*, *1*(3), 100027. https://doi.org/10.1016/j.rmal.2022.100027

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (PLS-SEM) using R: A workbook*. Springer Nature. https://doi.org/10.1007/978-3-030-80519-7

Hajiyeva, K. (2015). A corpus-based lexical analysis of subject-specific university textbooks for English majors. *Ampersand*, *2*, 136–144. https://doi.org/10.1016/j.amper.2015.10.001

Han, D., & Qian, D. D. (2024). Evaluating the roles of breadth and depth of aural vocabulary knowledge in listening comprehension of EFL learners: An investigation applying auditory measures. *System*, *120*, 103207. https://doi.org/10.1016/j.system.2023.103207

Hannibal Jensen, S. (2017). Gaming as an English language learning resource among young children in Denmark. *CALICO Journal*, *34*(1), 1–19. https://doi.org/10.1558/cj.29519

Harris, C. L. (1998). Psycholinguistic studies of entrenchment. In J. Koenig (Ed.), *Conceptual structures, language and discourse* (pp. 55–70). Center for the Study of Language and Information (CSLI).

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*(4), 839–872. https://doi.org/10.1111/lang.12353

He, X. (Stella), & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, *53*(2), 348–371. https://doi.org/10.1002/tesq.483

Heng, R., Pu, L., & Liu, X. (2023). The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1082228

Hennebry, M., Rogers, V., Macaro, E., & Murphy, V. (2017). Direct teaching of vocabulary after listening: Is it worth the effort and what method is best? *The Language Learning Journal*, *45*(3), 282–300. https://doi.org/10.1080/09571736.2013.849751

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Henriksen, B., & Danelund, L. (2015). Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English. In P. Pietilä, K. Doró, & R. Pípalová (Eds.),

Lexical issues in L2 writing (pp. 29–56). Cambridge Scholars Publishing.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* [Computer software]. https://doi.org/10.5281/zenodo.1212303

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

Housen, A., Janssens, S., & Pierrard, M. (2001). *Frans en Engels als vreemde talen in Vlaamse scholen: Een vergelijkend onderzoek*. VUB Press. https://lib.ugent.be/en/catalog/rug01:002022146

Hsu, W. (2009). College English textbooks for general purposes: A corpus-based analysis of lexical coverage. *Electronic Journal of Foreign Language Teaching*, *6*(1), 42–62.

Hu, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Hunt, A., & Beglar, D. (1998). Current research and practice in teaching vocabulary. *The Language Teacher*, *22*(1).

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development*, *9*(1–2), 67–84. https://doi.org/10.1080/01434632.1988.9994320

Iwaizumi, E., & Webb, S. (2022). Measuring L1 and L2 productive derivational knowledge: How many derivatives can L1 and L2 learners with differing vocabulary levels produce? *TESOL Quarterly*, *56*(1), 100–129. https://doi.org/10.1002/tesq.3035

Jamers, L. (2021). *A validation of the productive VocabLab test for English* [Unpublished master's thesis]. KU Leuven.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(Suppl. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of

phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843. https://doi.org/0278-7393/94/S3.00

Jin, Z., & Webb, S. (2025). Does writing words in notes contribute to vocabulary learning? *Language Teaching Research*, *29*(2), 560–587. https://doi.org/10.1177/13621688211062184

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, *7*(3), 700–715. https://doi.org/10.1002/tesj.238

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Kaatari, H., Larsson, T., Wang, Y., Acikara-Eickhoff, S., & Sundqvist, P. (2023). Exploring the effects of target-language extramural activities on students' written production. *Journal of Second Language Writing*, *62*, 101062. https://doi.org/10.1016/j.jslw.2023.101062

Karakoç, D., & Köse, G. D. (2017). The impact of vocabulary knowledge on reading, writing and proficiency scores of EFL learners. *Journal of Language and Linguistic Studies*, *13*(1), Article 1.

Kerekes, E. (2015). Using songs and lyrics in out-of-class learning. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 33–42). Routledge.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kisselev, O., Kopotev, M., & Vakhranev, A. (2025). Assessing lexical proficiency in Russian as a second language using indices of lexical sophistication, diversity, and fluency. *The Modern Language Journal*, *109*(2), 344–364. https://doi.org/10.1111/modl.12998

Knoch, U. (2021). Assessing writing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 236–253). Routledge.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, *4*(5), 900–913. https://doi.org/10.4304/jltr.4.5.900-913

Koizumi, R., & In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and Bayesian methods. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00618

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Krashen, S. (2004). *The power of reading: Insights from the research* (2nd ed.). Bloomsbury Publishing.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987.

Kremmel, B., Indrarathne, B., Kormos, J., & Suzuki, S. (2023). Unknown vocabulary density and reading comprehension: Replicating Hu and Nation (2000). *Language Learning*, *73*(4), 1127–1163. https://doi.org/10.1111/lang.12622

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Kuperman, V., Schroeder, S., & Gnetov, D. (2024). Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. *Journal of Memory and Language*, *135*, 104497. https://doi.org/10.1016/j.jml.2023.104497

Kuppens, A. H. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology*, *35*(1), 65–85. https://doi.org/10.1080/17439880903561876

Kurokawa, S., Hein, A. M., & Uchihara, T. (2024). Incidental vocabulary acquisition through captioned viewing: A meta-analysis. *Language Learning*, *Early View*. https://doi.org/10.1111/lang.12697

Kusyk, M., Arndt, H. L., Schwarz, M., Yibokou, K. S., Dressman, M., Sockett, G., & Toffoli, D. (2025). A scoping review of studies in informal second language learning: Trends in research published between 2000 and 2020. *System*, *130*, 103541. https://doi.org/10.1016/j.system.2024.103541

Kyle, K. (2020a). Lexis. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 332–344). Routledge.

Kyle, K. (2020b). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–476). Routledge.

Kyle, K. (2025). *Pylats* (Version 39) [Source code]. https://github.com/LCR-ADS-Lab/pylats

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., & Eguchi, M. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal*, *107*(2), 531–564. https://doi.org/10.1111/modl.12845

Kyle, K., & Eguchi, M. (2024). Evaluating NLP models with written and spoken L2 samples. *Research Methods in Applied Linguistics*, *3*(2), 100120. https://doi.org/10.1016/j.rmal.2024.100120

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, *46*(1), 278–299. https://doi.org/10.1017/S0272263123000402

Lai, C., & Gu, M. (2011). Self-regulated out-of-class language learning with technology. *Computer Assisted Language Learning*, *24*(4), 317–335. https://doi.org/10.1080/09588221.2011.568417

Langacker, R. W. (1987). *Foundations of cognitive grammar. Volume I: Theoretical prerequisites*. Stanford University Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Laurén & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In W. E. Rutherford, L. Eubank, L. Selinker, & M. Sharwood Smith (Eds.), *The current state of interlanguage* (pp. 265–272). John Benjamins.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review*, *59*(4), 567–587. https://doi.org/10.3138/cmlr.59.4.567

Laufer, B. (2005a). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*(1), 223–250. https://doi.org/10.1075/eurosla.5.11lau

Laufer, B. (2005b). Lexical Frequency Profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied Linguistics*, *26*(4), 582–588. https://doi.org/10.1093/applin/ami029

Laufer, B. (2014). Vocabulary in a second language: Selection, acquisition, and testing: A commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, *3*(2), 38–46. https://doi.org/10.7820/vli.v03.2.laufer

Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 351–368). Routledge.

Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729–741. https://doi.org/10.1111/modl.12431

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & McLean, S. (2016). Loanwords and Vocabulary Size Test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*(3), 202–217. https://doi.org/10.1080/15434303.2016.1210611

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Nation, I. S. P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? *EUROSLA Yearbook*, *1*(1), 7–28. https://doi.org/10.1075/eurosla.1.05lau

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge. https://doi.org/10.4324/9780203808184.ch10

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge.

Language Teaching Research, *19*(6), 687–711. https://doi.org/10.1177/1362168814559797

Laufer, B., & Vaisman, E. E. (2023). Out-of-classroom L2 vocabulary acquisition: The effects of digital activities and school vocabulary. *The Modern Language Journal*, *107*(4), 854–872. https://doi.org/10.1111/modl.12880

Lee, J. S. (2022). Evaluation of instruments for researching learners' LBC. In H. Reinders, C. Lai, & P. Sundqvist (Eds.), *The Routledge handbook of language learning and teaching beyond the classroom* (pp. 312–326). Routledge.

Lee, J. S., & Dressman, M. (2018). When IDLE hands make an English workshop: Informal digital learning of English and language proficiency. *TESOL Quarterly*, *52*(2), 435–445. https://doi.org/10.1002/tesq.422

Lemmouh, Z. (2010). *The relationship among vocabulary knowledge, academic achievement and the lexical richness in writing in Swedish university students of English* [Doctoral dissertation]. Stockholm University.

Leona, N. L., van Koert, M. J. H., van der Molen, M. W., Rispens, J. E., Tijms, J., & Snellings, P. (2021). Explaining individual differences in young English language learners' vocabulary knowledge: The role of Extramural English Exposure and motivation. *System*, *96*, 102402. https://doi.org/10.1016/j.system.2020.102402

Leow, R. P. (2019). Noticing hypothesis. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0086.pub2

Leow, R. P., Manchón, R. M., & Polio, C. (2022). Writing: Researching L2 writing as a site for learning in instructed settings. In L. Gurzynski-Weiss & Y. Kim (Eds.), *Instructed second language acquisition research methods* (pp. 305–328). John Benjamins. https://doi.org/10.1075/rmal.3.13leo

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer, *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 127–148). EuroSLA.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of Phrasal Verbs: The Case of Chinese Learners of English. *Language Learning*, *54*(2), 193–226. https://doi.org/10.1111/j.1467-9922.2004.00254.x

Libben, G. (2022). Lexical representations in language processing. In *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.398

Lin, H. F. (2023). Assessing the role of vocabulary knowledge in developing EFL learners' writing skills: Implications for intentional and incidental vocabulary learning. *Asian Journal of English Language Teaching*, *32*(1), 105–130.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. *The Canadian Modern Language Review*, *66*(3), 393–420. https://doi.org/10.3138/cmlr.66.3.393

Linnarud, M. (1986). *Lexis in composition. A performance analysis of Swedish learners' written English*. Gleerup.

Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). Routledge. https://doi.org/10.4324/9781315616797

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners* (1st ed.). Routledge.

López-Serrano, S., Roca de Larios, J., & Manchón, R. M. (2019). Language reflection fostered by individual L2 writing tasks: Developing a theoretically motivated and empirically based coding system. *Studies in Second Language Acquisition*, *41*(3), 503–527. https://doi.org/10.1017/S0272263119000275

Lorenzo-Dus, N. (2007). The best of both worlds? Combined methodological approaches to the assessment of vocabulary in oral proficiency interviews. In M. H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 220–233). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.015

Lorenzo-Dus, N., & Meara, P. (2004). Role-plays and the assessment of oral proficiency in Spanish. In R. Márquez Reiter & M. E. Placencia (Eds.), *Current trends in the pragmatics of Spanish* (pp. 79–98). John Benjamins.

Lowie, W., & Verspoor, M. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning*, *65*(1), 63–88. https://doi.org/10.1111/lang.12093

Lu, C., & Dang, T. N. Y. (2023). Effect of L2 exposure, length of study, and L2 proficiency on EFL learners' receptive knowledge of form–meaning connection and collocations of high-frequency words. *Language Teaching Research*, *Online First*, 1–27. https://doi.org/10.1177/13621688231155820

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lyrigkou, C. (2023). Informal second language learning and EFL learners' spoken use of discourse markers. In D. Toffoli, G. Sockett, & M. Kusyk (Eds.), *Language learning and leisure: Informal learning in the digital age* (pp. 91–120). De Gruyter. https://doi.org/10.1515/9783110752441-005

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift Für Literaturwissenschaft Und Linguistik*, *8*, 73–79.

Maie, R., Eguchi, M., & Uchihara, T. (2024). Arbitrary choices, arbitrary results: Three cases of multiverse analysis in L2 research. *Research Methods in Applied Linguistics*, *3*(2), 100124. https://doi.org/10.1016/j.rmal.2024.100124

Malmqvist, J., Hellberg, K., Möllås, G., Rose, R., & Shevlin, M. (2019). Conducting the pilot study: A neglected part of the research process? Methodological findings supporting the importance of piloting in qualitative research studies. *International Journal of Qualitative Methods*, *18*, 1609406919878341. https://doi.org/10.1177/1609406919878341

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230511804

Manchón, R. M. (2011). *Learning-to-write and writing-to-learn in an additional language*. John Benjamins. https://doi.org/10.1075/lllt.31

Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language*, *22*(1), 56–70. https://doi.org/10/articles/matsuoka.pdf

Maw, T. M. M., Clenton, J., & Higginbotham, G. (2022). Investigating whether a flemma count is a more distinctive measurement of lexical diversity. *Assessing Writing*, *53*, 100640. https://doi.org/10.1016/j.asw.2022.100640

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

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

Meara, P. (1988). Learning words in an L1 and an L2. *Polyglot*, *9*(3), 1–16.

Meara, P. (1990). A note on passive vocabulary. *Interlanguage Studies Bulletin (Utrecht)*, *6*(2), 150–154. https://doi.org/10.1177/026765839000600204

Meara, P. (2005a). Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics*, *26*(1), 32–47. https://doi.org/10.1093/applin/amh037

Meara, P. (2005b). *X_Lex, the Swansea Levels Test* (Version v2.05) [Computer software]. Lognostics.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospects*, *16*(3), 323–327.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society. Papers from the annual meeting of the British Association for Applied Linguistics* (pp. 80–87).

Meara, P., & Miralpeix, I. (2006). *Y_Lex, the Swansea Advanced Levels Test* (Version v2.05) [Computer software]. Lognostics.

Meara, P., & Miralpeix, I. (2016). P_Lex v3.00. In P. Meara & I. Miralpeix, *Tools for researching vocabulary*. Multilingual Matters.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge University Press.

Miller, M., & Hegelheimer, V. (2006). The Sims meet ESL: Incorporating authentic computer simulation games into the language classroom. *Interactive Technology and Smart Education*, *3*(4), 311–328. https://doi.org/10.1108/17415650680000070

Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*(2), 187–205. https://doi.org/10.1017/S0959269506002420

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

Milton, J. (2022). Vocabulary denial and the false god of structuralism in Ofsted's 2021 Curriculum Research Review for languages. *The Language Learning Journal*, *50*(2), 156–171. https://doi.org/10.1080/09571736.2022.2045680

Miralpeix, I. (2020). L1 and L2 vocabulary growth and size. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 189–206). Routledge.

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Mombaers, T., Gasse, R. V., & Maeyer, S. D. (2024). Learning from comPA(I)Ring exemplars: Enhancing genre knowledge of argumentative texts. *Journal of Writing Research*, *16*(1), 163–193. https://doi.org/10.17239/jowr-2024.16.01.06

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280–300. https://doi.org/10.1093/applin/amy056

Montero Perez, M., Sydorenko, T., & Huntley, L. (2024). Incidental vocabulary learning from watching audiovisual input. In M. F. Teng & B. L. Reynolds (Eds.), *Researching incidental vocabulary learning in a second language* (pp. 35–50). Routledge.

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292. https://doi.org/10.1002/rev3.3292

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

Muñoz, C. (2020). Boys like games and girls like movies: Age and gender differences in out-of-school contact with English. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *33*(1), 171–201. https://doi.org/10.1075/resla.18042.mun

Muñoz, C., & Cadierno, T. (2021). How do differences in exposure affect English language learning? A comparison of teenagers in two learning environments. *Studies in Second Language Learning and Teaching*, *11*(2), 185–212. http://doi.org/10.14746/ssllt.2021.11.2.2

Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners. *Language Learning*, *68*(4), 1076–1109. https://doi.org/10.1111/lang.12309

Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second*

Language Research, *39*(1), 13–37. https://doi.org/10.1177/02676583211015797

Myles, J. (2022). Second language writing and research: The writing process and error analysis in student texts. *Teaching English as a Second or Foreign Language - Electronic Journal*, *6*(2), 1–19.

Nahatame, S., & Uchida, S. (2025). Relative importance of lexical features in word processing during L2 English reading. *Studies in Second Language Acquisition*, *First View*, 1–24. https://doi.org/10.1017/S0272263125101137

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (2001). *Learning vocabulary in another language* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

Nation, I. S. P. (2012). *The Vocabulary Size Test*. Victoria University of Wellington. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009093873

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16. https://jalt-publications.org/content/index.php/jer/article/view/4

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

Nation, I. S. P., & Waring, R. (2024). Incidental vocabulary learning from extensive reading. In M. F. Teng & B. L. Reynolds (Eds.), *Researching incidental vocabulary learning in a second language* (pp. 19–34). Routledge.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary* (1st ed.). Heinle.

Neagu, T., Eklöf, H., Laukaityte, I., & Wedman, J. (2025). Reading, watching and gaming: Exploring the relationships between extramural English activities and academic L2 English reading comprehension in a Swedish university admissions test context.

Education Inquiry, *Online First*, 1–16. https://doi.org/10.1080/20004508.2025.2556572

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/S014271640707035X

Nguyen, C.-D. (2021). Lexical features of reading passages in English-language textbooks for Vietnamese high-school students: Do they foster both content and vocabulary gain? *RELC Journal*, *52*(3), 509–522. https://doi.org/10.1177/0033688219895045

Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual Vocabulary Size Test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Nicklin, C., Bailey, D., McLean, S., Kim, Y. A., Kang, H., & Vitta, J. P. (2025). Applying lexical sophistication models to wordlist development: A proof-of-concept study. *Research Methods in Applied Linguistics*, *4*(1), 100175. https://doi.org/10.1016/j.rmal.2024.100175

Niitemaa, M.-L. (2020). Informal acquisition of L2 English vocabulary. *Nordic Journal of Digital Literacy*, *15*(2), 86–105. https://doi.org/10.18261/issn.1891-943x-2020-02-02

Niu, L., Wang, X., Wallace, M. P., Pang, H., & Xu, Y. (2022). Digital learning of English as a foreign language among university students: How are approaches to learning linked to digital competence and technostress? *Journal of Computer Assisted Learning*, *38*(5), 1332–1346. https://doi.org/10.1111/jcal.12679

Noreillie, A.-S. (2019). *It's all about words. Three empirical studies into the role of lexical knowledge and use in French listening and speaking tasks* [Doctoral dissertation]. KU Leuven.

Noreillie, A.-S., Desmet, P., & Peters, E. (2020). Factors predicting low-intermediate French learners' vocabulary use in speaking tasks. *The Canadian Modern Language Review*, *76*(3), 194–217. https://doi.org/10.3138/cmlr-2019-0018

Noreillie, A.-S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages. *ITL - International Journal of Applied Linguistics*, *169*(1), 212–231. https://doi.org/10.1075/itl.00013.nor

Norris, J. M., & Ortega, L. (2001). Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning*, *51*(s1), 157–213. https://doi.org/10.1111/j.1467-1770.2001.tb00017.x

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 55–82). Brill. https://doi.org/10.1163/9789004248489_002

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237. https://doi.org/10.1177/026553229501200205

Olsson, E., & Sylvén, L. 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. https://doi.org/10.17011/apples/urn.201512234129

Onnis, L. (2012). The potential contribution of statistical learning to second language acquisition. In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (pp. 203–236). De Gruyter Mouton. https://doi.org/10.1515/9781934078242

OpenAI. (2024). *ChatGPT* (Version 2024) [Large Language Model].

Orosz, A. (2009). The growth of young learners' English vocabulary size. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 181–194). Multilingual Matters.

Ortega, L. (2009). Studying writing across EFL contexts: Looking back and moving forward. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 232–255). Multilingual Matters.

Ortega, L. (2013a). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, *63*(s1), 1–24. https://doi.org/10.1111/j.1467-9922.2012.00735.x

Ortega, L. (2013b). *Understanding second language acquisition*. Routledge.

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: Mesure fiable du niveau de compétence langagière. *The Canadian Modern Language Review*, *63*(1), 107–125. https://doi.org/10.3138/cmlr.63.1.107

Öztürk, M. (2022). Parts of speech distribution in the BNC-COCA word lists. *Language Education and Technology*, *2*(2), 128–140.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, *41*(4), 745–768. https://doi.org/10.1017/S0272263119000020

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition *from* and *while* reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*(1), 97–130. https://doi.org/10.1017/S0272263115000224

Pellicer-Sánchez, A., & Montero Perez, M. (2024). Eye-tracking in vocabulary research: Introduction to the special issue. *Research Methods in Applied Linguistics*, *3*(1), 100095. https://doi.org/10.1016/j.rmal.2024.100095

Pellicer-Sánchez, A., Siyanova-Chanturia, A., & Parente, F. (2022). The effect of frequency of exposure on the processing and learning of collocations: A comparison of first and second language readers' eye movements. *Applied Psycholinguistics*, *43*(3), 727–756. https://doi.org/10.1017/S014271642200011X

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins.

Peters, E. (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. https://doi.org/10.1075/itl.00010.pet

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, *63*, 134–148. https://doi.org/10.1016/J.SYSTEM.2016.10.002

Peters, E., Noreillie, A.-S., Heylen, K., Bulté, B., & Desmet, P. (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. https://doi.org/10.1111/lang.12351

Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests: The development of an English and French vocabulary test. *ITL - International Journal of Applied Linguistics*, *170*(1), 53–78. https://doi.org/10.1075/ITL.17029.PET

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

Pfenninger, S. E. (2020). The dynamic multicausality of age of first bilingual language exposure: Evidence from a longitudinal content and language integrated learning study with dense time

serial measurements. *The Modern Language Journal*, *104*(3), 662–686. https://doi.org/10.1111/modl.12666

Pfenninger, S. E., & Singleton, D. (2016). Age of onset, socio-affect and cross-linguistic influence: A long-term classroom study. *Vigo International Journal of Applied Linguistics*, *13*, 147–179.

Pfenninger, S. E., & Singleton, D. (2019). Starting age overshadowed: The primacy of differential environmental and family support effects on second language attainment in an instructional context. *Language Learning*, *69*(S1), 207–234. https://doi.org/10.1111/lang.12318

Pinchbeck, G. G., Brown, D., Mclean, S., & Kramer, B. (2022). Validating word lists that represent learner knowledge in EFL contexts: The impact of the definition of word and the choice of source corpora. *System*, *106*, 102771. https://doi.org/10.1016/j.system.2022.102771

Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, *8*(1), 75. https://doi.org/10.3390/languages8010075

Prophète, E., Sundqvist, P., & Peters, E. (2024, July 5). *Learning English beyond the classroom: Young EFL learners' grammaticality judgement and writing proficiency prior to formal English instruction* [Conference presentation]. EuroSLA, Montpellier, France.

Puimège, E. (2022). *An empirical investigation into the acquisition and processing of L2 formulaic sequences in meaningful input* [Doctoral dissertation]. KU Leuven.

Puimège, E. (2024). Learning L2 formulaic sequences from meaning-focused activities. *ITL - International Journal of Applied Linguistics*, *175*(2), 163–186. https://doi.org/10.1075/itl.23014.pui

Puimège, E., Montero Perez, M., & Peters, E. (2024). The effects of typographic enhancement on L2 collocation processing and learning from reading: An eye-tracking study. *Applied Linguistics*, *45*(1), 88–110. https://doi.org/10.1093/applin/amad003

Puimège, E., & Peters, E. (2019a). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, *69*(4), 943–977. https://doi.org/10.1111/lang.12364

Puimège, E., & Peters, E. (2019b). Learning L2 vocabulary from audiovisual input: An exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, *47*(4), 424–438. https://doi.org/10.1080/09571736.2019.1638630

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, *47*(4), 479–496. https://doi.org/10.1080/09571736.2019.1616806

Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' TV viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, *42*(3), 551–575. https://doi.org/10.1017/S0272263120000042

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, *56*(2), 282–308. https://doi.org/10.3138/cmlr.56.2.282

Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge.

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ranalli, J. (2008). Learning English with The Sims: Exploiting authentic computer simulation games for L2 learning. *Computer Assisted Language Learning*, *21*(5), 441–455. https://doi.org/10.1080/09588220802447859

Ray, S., Danks, N. P., & Valdez, A. C. (2024). *seminr: Building and estimating Structural Equation Models* [Computer software]. https://CRAN.R-project.org/package=seminr

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. https://doi.org/10.1177/026553229301000308

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Read, J. (2020). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

Rebuschat, P., & Williams, J. (2012). Introduction: Statistical learning and language acquisition. In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (pp. 1–12). De Gruyter Mouton. https://doi.org/10.1515/9781934078242

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–210). Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.017

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-

gaze behaviours in L2 writing. *Journal of Second Language Writing*, *58*, 100927. https://doi.org/10.1016/j.jslw.2022.100927

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. https://doi.org/10.2307/3585941

Richards, J. C., & Schmidt, R. W. (2002). *Longman dictionary of language teaching and applied linguistics* (3rd ed.). Routledge. https://doi.org/10.4324/9781315833835

Rød, A. J., & Calafato, R. (2025). Bridging extramural English and formal instruction. *ELT Journal*, *79*(3), 363–371. https://doi.org/10.1093/elt/ccaf026

Rodgers, M. P. H. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions* [Doctoral dissertation]. Victoria University of Wellington. https://doi.org/10.26686/wgtn.17005078

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689–717. https://doi.org/10.5054/tq.2011.268062

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249–284. https://doi.org/10.1016/S0010-0277(97)00027-9

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research*, *21*(1), 12–30. https://doi.org/10.1177/1362168815586083

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmitt, D. (2020). Can writing facilitate the development of a richer vocabulary? Advancing research agendas. In R. M. Manchón (Ed.), *Writing and language learning: Advancing research agendas* (pp. 357–380). John Benjamins.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2010a). Key issues in teaching and learning vocabulary. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 28–40). Multilingual Matters.

Schmitt, N. (2010b). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274. https://doi.org/10.1017/S0261444819000053

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145–171. https://doi.org/10.2307/3588328

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Schoonen, R. (2011). How language ability is assessed. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 701–716). Routledge.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1–22). Brill. https://doi.org/10.1163/9789004248489_002

Schoonen, R., Gelderen, A. van, Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165–202. https://doi.org/10.1111/1467-9922.00213

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211–236. https://doi.org/10.1177/0265532207086782

Schurz, A., & Sundqvist, P. (2022). Connecting Extramural English with ELT: Teacher reports from Austria, Finland, France, and Sweden.

Applied Linguistics, *43*(5), 934–957. https://doi.org/10.1093/applin/amac013

Schwarz, M. (2020). *Beyond the walls: A mixed methods study of teenagers' extramural English practices and their vocabulary knowledge* [Doctoral dissertation]. University of Vienna. https://doi.org/10.25365/thesis.63632

Schwarz, M., & Ghamarian-Krenn, K. (2024). Extramural engagement with English and vocabulary learning outcomes: A comparative account of research on Viennese teenagers and university students. In U. Smit & S. Reichl (Eds.), *#YouthMediaLife & Friends: Interdisciplinary research into young people's mediatised lifeworlds* (pp. 43–67). V&R unipress. https://doi.org/10.14220/9783737016391.43

Shapiro, B. J. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 248–251. https://doi.org/10.1016/S0022-5371(69)80070-8

Shukla, M., Gervain, J., Mehler, J., & Nespor, M. (2012). Linguistic constraints on statistical learning in early language acquisition. In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (pp. 171–202). De Gruyter Mouton. https://doi.org/10.1515/9781934078242

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, *53*, 148–160. https://doi.org/10.1016/j.system.2015.07.003

Snoder, P., & Laufer, B. (2022). EFL learners' receptive knowledge of derived words: The case of Swedish adolescents. *TESOL Quarterly*, *56*(4), 1242–1265. https://doi.org/10.1002/tesq.3101

Sockett, G. (2014). *The online informal learning of English*. Palgrave Macmillan.

Sockett, G. (2022). Learning beyond the classroom and autonomy. In H. Reinders, C. Lai, & P. Sundqvist (Eds.), *The Routledge handbook of language learning and teaching beyond the classroom* (pp. 67–80). Routledge. https://doi.org/10.4324/9781003048169-7

Sockett, G. (2025, February 7). *Online informal learning of English* [Interview]. Talking about languages, University of Vigo. https://tv.uvigo.es/video/67b5c1ce21ce6d813f15f102

Sockett, G., & Kusyk, M. (2015). Online informal learning of English: Frequency effects in the uptake of chunks of language from participation in web-based activities. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 153–178). De Gruyter. https://doi.org/10.1515/9783110378528-009

Sockett, G., & Toffoli, D. (2012). Beyond learner autonomy: A dynamic systems view of the informal learning of English in virtual online

communities. *Procedia - Social and Behavioral Sciences*, *34*, 212–215. https://doi.org/10.1016/j.sbspro.2012.02.042

Song, Y., & Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: Effects of strategy instruction on college students' argumentative essays. *Reading and Writing*, *26*(1), 67–90. https://doi.org/10.1007/s11145-012-9381-8

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, *41*(1), 89–108. https://doi.org/10.1177/02655322231162853

Stryckmans, A. (2025). *Writing proficiency in English and French among L2 learners in secondary school: A lexical error analysis* [Unpublished bachelor's thesis]. KU Leuven.

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 102279. https://doi.org/10.1016/j.system.2020.102279

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary* [Doctoral dissertation]. Karlstad University.

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology*, *23*(1), 87–113. https://doi.org/10125/44674

Sundqvist, P. (2024). Extramural English as an individual difference variable in L2 research: Methodology matters. *Annual Review of Applied Linguistics, 44*, 79–91. https://doi.org/10.1017/S0267190524000072

Sundqvist, P., & Sylvén, L. K. (2014). Language-related computer use: Focus on young L2 English learners in Sweden. *ReCALL*, *26*(1), 3–20. https://doi.org/10.1017/S0958344013000232

Sundqvist, P., & Uztosun, M. S. (2023). Extramural English in Scandinavia and Asia: Scale development, learner engagement, and perceived

speaking ability. *TESOL Quarterly, 58*(4), 1638–1665. https://doi.org/10.1002/tesq.3296

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System*, *51*, 65–76. https://doi.org/10.1016/j.system.2015.04.001

Sung, H., Cho, S., & Kyle, K. (2024). An empirical evaluation of lexical diversity indices in L2 Korean writing assessment. *Language Assessment Quarterly*, *21*(2), 159–180. https://doi.org/10.1080/15434303.2024.2311728

Sung, H., Wolf, M. K., Suhan, M., & Kyle, K. (2025). Lexical richness in young English learners' writing: A focus on opinion and listen-write task types. *Assessing Writing*, *66*, 100975. https://doi.org/10.1016/j.asw.2025.100975

Swain, M. (1985). Communicative competene: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 471–483). Routledge.

Sylvén, L. K. (2006). How is extramural exposure to English among Swedish school students used in the CLIL classroom? *Vienna English Working Papers*, *15*(3), 47–53.

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, *24*(3), 302–321. https://doi.org/10.1017/S095834401200016X

Tahmouresi, S., & Papi, M. (2021). Future selves, enjoyment and anxiety as predictors of L2 writing achievement. *Journal of Second Language Writing*, *53*, 100837. https://doi.org/10.1016/j.jslw.2021.100837

Takala, S. (1985). Estimating students' vocabulary sizes in foreign language teaching. In V. Kohonen, H. Von Hessen, & C. Klein-Braley (Eds.), *Practice and problems in language testing* (pp. 157–165). AfinLa.

Tantucci, V., Bottini, R., & Wang, A. (2025). Do examiners and test-takers imitate each other? Dialogic resonance in second language testing. *Applied Linguistics*, *Advance online publication*, amaf044. https://doi.org/10.1093/applin/amaf044

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tracy-Ventura, N. (2017). Combining corpora and experimental data to investigate language learning during residence abroad: A study of lexical sophistication. *System*, *71*, 35–45. https://doi.org/10.1016/j.system.2017.09.022

Tracy-Ventura, N., Huensch, A., & Mitchell, R. (2021). Understanding the long-term evolution of L2 lexical diversity: The contribution of a longitudinal learner corpus. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 148–171). Cambridge University Press. https://doi.org/10.1017/9781108674577.008

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). John Benjamins. https://doi.org/10.1075/sibil.47.05ch3

Treffers-Daller, J., Daller, M. H., Malvern, D., Richards, B., Meara, P., & Milton, J. (2008). Introduction: Special issue on knowledge and use of the lexicon in French as a second language. *Journal of French Language Studies*, *18*(3), 269–276. https://doi.org/10.1017/S095926950800344X

Treffers-Daller, J., & Korybski, T. (2015). Using lexical diversity measures to operationalize language dominance in bilinguals. In C. Silva-Corvalán & J. Treffers-Daller (Eds.), *Language dominance in bilinguals: Issues of measurement and operationalization* (pp. 106–133). Cambridge University Press. https://doi.org/10.1017/CBO9781107375345.006

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Tschirner, E. (2004). Breadth of vocabulary and advanced English study: An empirical investigation. *Electronic Journal of Foreign Language Teaching*, *1*(1), 26–38.

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, *24*(4), 540–556. https://doi.org/10.1177/1362168818799371

Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*(1), 64–75. https://doi.org/10.1080/09571736.2016.1191527

Uchihara, T., Saito, K., & Clenton, J. (2020). Re-examining the relationship between productive vocabulary and second language oral ability. In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills* (pp. 146–165). Routledge.

Unsworth, S., Persson, L., Prins, T., & De Bot, K. (2014). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, *36*(5), 527–548. https://doi.org/10.1093/applin/amt052

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perrin & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443–452). Cambridge University Press.

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

Van Dale. (n.d.). https://vowb-vandale-nl.kuleuven.e-bronnen.be/

Van den Branden, K. (2022). *How to teach an additional language: To task or not to task?* John Benjamins. https://doi.org/10.1075/tblt.15

Van Parys, A., De Wilde, V., Macken, L., & Montero Perez, M. (2024). Vocabulary of reading materials in English and French L2 textbooks: A cross-lingual corpus study. *System*, *124*, 103396. https://doi.org/10.1016/j.system.2024.103396

Van Parys, A., De Wilde, V., Macken, L., & Montero Perez, M. (2025). LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research, Online First*. https://doi.org/10.1177/13621688251352259

Van Steendam, E., Vandermeulen, N., & Rijlaarsdam, G. (2025, August 28). *Predicting text quality and epistemic experience in source-based writing* [Conference presentation]. EARLI, Graz, Austria.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074

Vandeweerd, N., Housen, A., & Paquot, M. (2023). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition, 45*(4), 787–811. https://doi.org/10.1017/S0272263122000389

Vandommele, G., Van den Branden, K., Van Gorp, K., & De Maeyer, S. (2017). In-school and out-of-school multimodal writing as an L2 writing resource for beginner learners of Dutch. *Journal of Second Language Writing*, *36*, 23–36. https://doi.org/10.1016/j.jslw.2017.05.010

Vilkaitė-Lozdienė, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge.

Vrancx, S. (2016). *Franse woordenschattoetsen: Een vergelijking tussen een eentalige en een tweetalige versie* [Unpublished master's thesis]. KU Leuven.

Vu, D. V., & Peters, E. (2022). Learning vocabulary from reading-only, reading-while-listening, and reading with textual input

enhancement: Insights from Vietnamese EFL learners. *RELC Journal*, *53*(1), 85–100. https://doi.org/10.1177/0033688220911485

Wagner, J. (2015). Designing for language learning in the wild: Creating social infrastructures for second language learning. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 75–102). De Gruyter Mouton.

Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*(1), 5–44. https://doi.org/10.1111/lang.12424

Wang, A., & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*, *72*(3), 765–805. https://doi.org/10.1111/lang.12495

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata, The Occasional Papers at Notre Dame Seishin University*, *1*, 53–68.

Warnby, M. (2022). Receptive academic vocabulary knowledge and extramural English involvement – is there a correlation? *ITL - International Journal of Applied Linguistics*, *173*(1), 120–152. https://doi.org/10.1075/itl.21021.war

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360–376. https://doi.org/10.1177/0033688209343854

Webb, S. (2013). Depth of vocabulary knowledge. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing. https://doi.org/10.1002/9781405198431.wbeal1325

Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159–168). Routledge.

Webb, S. (Ed.). (2020). *The Routledge handbook of vocabulary studies* (1st ed.). Routledge. https://doi.org/10.4324/9780429291586

Webb, S. (2021a). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, *33*(2), 278–293. https://doi.org/10.64152/10125/67407

Webb, S. (2021b). The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, *43*(5), 941–949. https://doi.org/10.1017/S0272263121000784

Webb, S., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126. https://doi.org/10.1177/0033688212439367

Webb, S., Pellicer-Sánchez, A., & Wang, A. (2025). Lexical coverage and reading comprehension revisited. *Reading in a Foreign Language*, *37*(1), 1–21. https://doi.org/10.64152/10125/67517

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., Sasao, Y., & Ballance, O. (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. https://doi.org/10.1075/itl.168.1.02web

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, *56*(2), 161–180. https://doi.org/10.1017/S0261444822000507

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715–738. https://doi.org/10.1111/modl.12671

Weigle, S. C. (2010). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Werner, V., & Tegge, F. (Eds.). (2021). *Pop culture in language education: Theory, research, practice*. Routledge.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

Whaley, C. P. (1978). Word–nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154. https://doi.org/10.1016/S0022-5371(78)90110-X

Woods, K., Hashimoto, B., & Brown, E. K. (2023). A multi-measure approach for lexical diversity in writing assessments: Considerations in measurement and timing. *Assessing Writing*, *55*, 100688. https://doi.org/10.1016/j.asw.2022.100688

Wouters, M., Bollansée, L., Prophète, E., & Peters, E. (2024). The relationship between extramural English and learners' listening comprehension, reading comprehension, motivation, and

anxiety. *Vigo International Journal of Applied Linguistics*, *21*, 165–193. https://doi.org/10.35869/vial.v0i21.4570

Wulff, S. (2021). Krashen's claims through a usage-based lens. *Foreign Language Annals*, *54*(2), 306–310.

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371–386). Routledge.

Yanagisawa, A., & Webb, S. (2022). Involvement Load Hypothesis Plus: Creating an improved predictive model of incidental vocabulary learning. *Studies in Second Language Acquisition*, *44*(5), 1279–1308. https://doi.org/10.1017/S0272263121000577

Yang, Y., Sun, Y., Chang, P., & Li, Y. (2019). Exploring the relationship between language aptitude, vocabulary size, and EFL graduate students' L2 writing performance. *TESOL Quarterly*, *53*(3), 845–856.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007

Yoon, H.-J., & Tabari, M. A. (2023). Authorial voice in source-based and opinion-based argumentative writing: Patterns of voice across task types and proficiency levels. *Journal of English for Academic Purposes*, *62*, 101228. https://doi.org/10.1016/j.jeap.2023.101228

Yu, S., Liu, C., & Zhang, L. (2023). Understanding L2 writers' lived experiences of informal writing: A phenomenological approach. *Journal of Second Language Writing*, *60*, 100979. https://doi.org/10.1016/j.jslw.2023.100979

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595. https://doi.org/10.1017/S0272263105050254

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhang, R., Zou, D., Cheng, G., Xie, H., Wang, F. L., & Au, O. T. S. (2021). Target languages, types of activities, engagement, and effectiveness of extramural language learning. *PLOS ONE*, *16*(6), e0253431. https://doi.org/10.1371/journal.pone.0253431

Zhang, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics*, *32*(3), 371–396. https://doi.org/10.1111/ijal.12420

Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*, *35*(3), 283–304. https://doi.org/10.1093/applin/amt014

Appendices

A1. Productive vocabulary tests

A1.1. Productive VocabLab test – English

Je krijgt nu een reeks **Engelse zinnen** waarin telkens een woord is weggelaten.

👉 **Vul het ontbrekende woord in**.

- Je krijgt altijd de eerste letter(s) van het woord. De streepjes geven aan hoeveel letters je nog moet aanvullen.
- Achteraan elke zin vind je een Nederlandse vertaling van het woord dat we zoeken.
- Schrijf steeds het volledige woord op.

⏰ Je krijgt hiervoor in totaal **30 minuten**.

Voorbeeld:
He loves doing wrong and does not follow any of the r_ _ _ _. *(regels)*
Antwoord: rules

Deel 1

1.	This might s_ _ _ like a very difficult problem, but it is not. *(lijken)*	1.
2.	I don't l_ _ _ people who ask questions all the time. *(leuk vinden)*	2.
3.	His head is too big for his b_ _ _. *(lichaam)*	3.
4.	When you begin to r_ _ _ that book, let me know if you enjoy it. *(lezen)*	4.
5.	There is a very high l_ _ _ _ of crime in this area. *(niveau)*	5.
6.	I have a car, so if you want me to, I can d_ _ _ _ you home. *(rijden)*	6.
7.	You can see the beauty of nature if you look right out of your w_ _ _ _ _. *(raam)*	7.
8.	What time does your plane a_ _ _ _ _ here? Maybe we can pick you up from the airport. *(aankomen)*	8.
9.	We've only got time for one fi_ _ _ question. *(laatste)*	9.

10. The product is undergoing a last tr_ _ _. If the tests are good, it will be on sale next week. *(test)* 10.
11. As she opened the box, she hoped it would c_ _ _ _ _ _ their love letters. *(bevatten)* 11.
12. Are you sitting comfortably in your ch_ _ _? *(stoel)* 12.
13. The most important c_ _ _ _ of accidents is that cars go too fast. *(oorzaak)* 13.
14. I bought this car from someone else, but the previous o_ _ _ _ obviously forgot to clean it. *(eigenaar)* 14.
15. You can tell me if I am wrong, but I a_ _ _ _ _ that everyone here has an email address? *(veronderstel)* 15.
16. The images were sold to the p_ _ _ _ and were published in every single newspaper. *(pers)* 16.
17. Who is the e_ _ _ _ _ of the New York Times? *(redacteur)* 17.
18. She was travelling at the sp_ _ _ of 90 miles per hour! *(snelheid)* 18.
19. The ch_ _ _ officer leads this department. *(hoofd-)* 19.
20. Be careful with fire, it can be very d_ _ _ _ _ _ _ _! *(gevaarlijk)* 20.
21. When you have finished, please put the books back where they b_ _ _ _ _. *(thuishoren)* 21.
22. This picture shows Ernest Hemingway, sitting at his d_ _ _ where he wrote most of his books. *(bureautafel)* 22.
23. The United States is going to sh_ _ _ focus to Asia. They are no longer interested in Europe. *(verschuiven)* 23.
24. Could you please m_ _ _ the route through the town on our map with this pencil? *(aanduiden)* 24.
25. When the city became under attack, a lot of young men came forward, willing to d_ _ _ _ _ all of us. *(verdedigen)* 25.
26. Wait, they will show the goal again in slow m_ _ _ _ _. *(beweging)* 26.
27. Unfortunately, there is only little research on this very interesting to_ _ _. *(onderwerp)* 27.
28. They were very close when they were children and today, they still have a very strong b_ _ _. *(band)* 28.

29. Chelsea Football Club are top of the l_ _ _ _ _ again. They are the true champions. *(divisie)* 29.

30. Our dog eats a v_ _ _ amount of food each day. No wonder he is so big! *(groot)* 30.

Deel 2

31. You can use capitals or highlight to e_ _ _ _ _ _ _ _ a word in this text. *(benadrukken)* 31.

32. Let's hope she doesn't a_ _ _ _ _ him of lying during the job interview. *(beschuldigen)* 32.

33. I would have done the job myself, but I do not have the p_ _ _ _ _ tools. *(geschikte)* 33.

34. I believe it is a great h_ _ _ _ _ to be here with you tonight. *(eer)* 34.

35. Bob is a recent mathematics gr_ _ _ _ _ _. He got his degree from the University of Oxford. *(afgestudeerde)* 35.

36. The t_ _ _ _ _ _ _ _ _ between winter and spring takes only a few weeks. *(overgang)* 36.

37. You should l_ _ _ your car or it will get stolen. *(afsluiten)* 37.

38. There's a lovely s_ _ _ _ _ that flows through their garden. *(beek)* 38.

39. We need to t_ _ _ _ with Eastern Europe if we want to remain competitive. *(handel drijven)* 39.

40. Telephone signals usually come through metal w_ _ _. *(draad)* 40.

41. The Bible and the Koran are important h_ _ _ books. *(heilige)* 41.

42. It is only a two-hour train j_ _ _ _ _ _ from Brussels to London. *(reis)* 42.

43. He had eaten so much that he had to loosen his b_ _ _. *(broekriem)* 43.

44. The doctors expect that she will make a full r_ _ _ _ _ _ _ from the operation soon. *(herstel)* 44.

45. She had been walking in the rain and was completely w_ _. *(nat)* 45.

46. Your job is to a_ _ _ _ _ the head chef and help him in the kitchen. *(bijstaan)*	46.
47. She won an a_ _ _ _ for her performance in the Disney movie. *(prijs)*	47.
48. Be careful when you say the p_ _ _ _ _ "how hard can it be?". *(uitdrukking)*	48.
49. I once met a r_ _ _ _ _ _ who had tried to cross our border to start a new life here. *(vluchteling)*	49.
50. I could not hear myself think anymore. The neighbours continued to y_ _ _ mean things at each other for hours. *(schreeuwen)*	50.
51. You should clean the garage, s_ _ _ _ the floor and do the washing-up. *(schoonvegen)*	51.
52. After the t_ _ _ _ _ _ _ _ of the main witness, everyone believed that she was guilty. *(getuigenis)*	52.
53. I always use a piece of t_ _ _ _ _ to clean my sunglasses. *(zakdoekje)*	53.
54. Bonus payments will give people an i_ _ _ _ _ _ _ _ to work harder. *(stimulans, drijfveer)*	54.
55. The teacher had to d_ _ _ _ _ _ the class early because she had a meeting. *(wegsturen)*	55.
56. If you sp_ _ any mistakes in this text, please mark them with a red pen. *(opmerkt)*	56.
57. Can you help me out? I cannot get the plates on the top s_ _ _ _ of the cupboard. *(kastplank)*	57.
58. Tourists are v_ _ _ _ _ _ _ _ _ to attack, because they do not know which areas of the city they should avoid. *(kwetsbaar)*	58.
59. During the war most people decided to f_ _ _ the country and start a new life here. *(ontvluchten)*	59.
60. I cannot enjoy this, I am simply too s_ _ _ _ _ to go bungee jumping. *(bang)*	60.

Dank je wel om deze woordenschattaak te maken!

A1.2. Productive VocabLab test – French

Je krijgt nu een reeks **Franse zinnen** waarin telkens een woord is weggelaten.

👉 **Vul het ontbrekende woord in**.

- Je krijgt altijd de eerste letter(s) van het woord. De streepjes geven aan hoeveel letters je nog moet aanvullen.
- Achteraan elke zin vind je een Nederlandse vertaling van het woord dat we zoeken.
- Schrijf steeds het volledige woord op.

⏰ Je krijgt hiervoor in totaal **30 minuten**.

Voorbeeld:
La nuit, je fais des r_ _ _ _ extraordinaires où je passe des aventures incroyables. *(dromen)*
Antwoord: rêves

Deel 1

1.	Philippe est ce_ _ _ _ _ que les résultats des deux matchs suivants seront décisifs. *(zeker)*	1.
2.	Mon grand-père est m_ _ _ à la suite d'une grave maladie. *(dood)*	2.
3.	Le maire a interdit la manifestation de ce mouvement d'extrême droite. Ainsi, il voulait em_ _ _ _ _ _ des problèmes entre des manifestants de gauche et de droite. *(verhinderen)*	3.
4.	Dans une démocratie, c'est le p_ _ _ _ _ qui doit choisir son gouvernement. *(volk)*	4.
5.	Elle avait l'air sympa avec ce b_ _ _ pantalon. *(mooi)*	5.
6.	La p_ _ _ _ de la Bastille a été un évènement important dans l'histoire de la France. *(verovering)*	6.
7.	Olivier veut q_ _ _ _ _ _ sa femme pour une autre. *(verlaten)*	7.
8.	Le pôle Sud, c'est l'e_ _ _ _ _ _ le plus froid de la planète. *(plaats)*	8.
9.	Il n'a pas pu p_ _ _ _ _ _ _ _ _ à Roland Garros à cause d'une blessure. *(deelnemen)*	9.
10.	La l_ _ _ _ contre la criminalité est une priorité du nouveau maire. *(strijd)*	10.

11.	Cette entreprise veut c_ _ _ _ _ _ _ _ _ la tour la plus haute d'Europe. *(bouwen)*	11.
12.	L'a_ _ _ _ _ _ de pouvoir acheter des livres et dvd en ligne, c'est qu'il ne faut plus se déplacer. *(voordeel)*	12.
13.	Je peux te demander une f_ _ _ _ _ ? Je cherche quelqu'un pour traduire ce texte en anglais. *(gunst)*	13.
14.	Tu peux m'aider ? Ce carton est trop l_ _ _ _ pour être porté par une seule personne. *(zwaar)*	14.
15.	Le chasseur avait touché le lapin : il y avait du s_ _ _ partout. *(bloed)*	15.
16.	Il p_ _ _ _ _ _ avoir serré la main au Président des États-Unis, mais je ne le crois pas. *(beweert)*	16.
17.	Je n'ai plus confiance en lui parce que la dernière fois, il n'a pas tenu sa p_ _ _ _ _ _ _. Il ne fait jamais ce qu'il a dit. *(belofte)*	17.
18.	Le tigre est devenu un animal r_ _ _. *(zeldzaam)*	18.
19.	Dans la soirée du 20 juin, la rock star assistera à une cérémonie à Paris, et puis le l_ _ _ _ _ _ _, elle rendra visite à des enfants malades. *(volgende dag)*	19.
20.	Juste après la n_ _ _ _ _ _ _ _ de sa fille, il a mis les premières photos sur Instagram. *(geboorte)*	20.
21.	Sa f_ _ en Dieu aide le croyant à surmonter toutes les difficultés. *(geloof)*	21.
22.	J'ai froid. Tu peux mettre du b_ _ _ sur le feu ? *(hout)*	22.
23.	Ça se voit qu'il est heureux. Il a toujours un grand s_ _ _ _ _ _. *(glimlach)*	23.
24.	Il a beaucoup étudié, mais je ne sais pas si ce sera s_ _ _ _ _ _ _ pour réussir. *(voldoende)*	24.
25.	- Tu as lu «La carte et le territoire» ? - Ah oui, Michel Houellebecq est un grand é_ _ _ _ _ _ _. *(schrijver)*	25.
26.	Un mode de vie plus s_ _ _ permettrait de prévenir beaucoup de problèmes physiques. *(gezond)*	26.
27.	Donne-lui une b_ _ _ _ de bonbons afin de le féliciter pour sa promotion. *(doos)*	27.
28.	Son appartement témoigne d'un bon g_ _ _. J'aime bien les couleurs fraîches et les meubles design. *(smaak)*	28.

29. Il a découvert une pièce de deux euros dans la p_ _ _ _ de son pantalon. *(zak)* 29.

30. L'âge agit sur le c_ _ _ _ _ _. Voilà pourquoi il faut rester actif aussi au niveau intellectuel : lire des livres, apprendre une nouvelle langue, faire des sudokus, etc. *(hersenen)* 30.

Deel 2

31. Il fait trop chaud pour se mettre au soleil. Il faut s'installer à l'o_ _ _ _. *(schaduw)* 31.

32. Après son s_ _ _ _ _ aux soins intensifs, le prince poursuit sa revalidation dans un autre service. *(verblijf)* 32.

33. L'expression d'une c_ _ _ _ _ _ _ _ _ _ religieuse appartient à la sphère privée et tout le monde devrait avoir le droit de s'exprimer librement. *(overtuiging)* 33.

34. Nous vous offrons un lit d_ _ _ et confortable et un petit-déjeuner buffet soigné. *(zacht)* 34.

35. Depuis ses origines, notre produit fonde sa légitimité sur son expertise du soin de la p_ _ _ sensible car rien n'est plus fragile que la p_ _ _ du bébé qui ne tolère pas la moindre allergie. *(huid)* 35.

36. On a dû couper cette b_ _ _ _ _ _ du grand arbre pour éviter qu'elle ne tombe sur notre maison. *(tak)* 36.

37. Tu as aussi entendu la r_ _ _ _ _ selon laquelle le président aurait une affaire avec une actrice ? *(gerucht)* 37.

38. Le représentant de cette organisation a demandé de mieux r_ _ _ _ _ _ _ les bénéfices entre les populations africaines. *(verdelen)* 38.

39. Vraiment ? Elle est e_ _ _ _ _ _ _ depuis cinq mois ? Je ne savais pas qu'elle attendait un enfant. *(zwanger)* 39.

40. Au cours de la journée, le temps devrait rester s_ _ dans le sud du pays. Il y aura éventuellement de la pluie dans le nord. *(droog)* 40.

41. Après avoir infecté les PC, les virus peuvent maintenant se r_ _ _ _ _ _ _ aussi sur les smartphones et les tablettes. *(verspreiden)* 41.

42. J'adore les pâtes avec de l'h_ _ _ _ d'olive. *(olie)* 42.

43. En voyant son cadavre, elle a poussé un c_ _ _ de terreur. *(schreeuw)* 43.

44. Pourquoi les chiens veulent-ils toujours c_ _ _ _ _ _ des trous ? *(graven)* 44.

45. Avant de passer à la seconde c_ _ _ _ _, il faut être sûr que la peinture est sèche. *(laag)* 45.

46. L'é_ _ _ _ _ relie le bras avec le reste du corps. *(schouder)* 46.

47. Je vois de l'eau c_ _ _ _ _ le long du mur de ma chambre ! Qu'est-ce qu'il faut faire ? *(stromen)* 47.

48. Si vous avez encore des questions, adressez-vous à l'a_ _ _ _ _ _ où on vous aidera volontiers. *(onthaal)* 48.

49. Je voudrais savoir qui est le cycliste le plus rapide et qui est le plus l_ _ _. *(traag)* 49.

50. Nous sommes de plus en plus nombreux à être attentifs à la p_ _ _ _ _ _ _ _ _ des produits et à préférer ceux de chez nous. *(herkomst)* 50.

51. J'ai appris que vous serez bientôt père et j'en suis r_ _ _. *(verheugd)* 51.

52. Google a réalisé une nouvelle a_ _ _ _ _ _ _ _ _ _ en achetant Boston Dynamics. *(aanwinst)* 52.

53. Il nous a appris que le rêve et la réalité sont deux concepts dis_ _ _ _ _ _. *(verschillend)* 53.

54. Vous prenez l'escalier jusqu'au deuxième étage. Ensuite, vous arrivez dans le c_ _ _ _ _ _ et vous frappez à la dernière porte sur votre droite où se trouve son bureau. *(gang)* 54.

55. Quel est votre l_ _ _ _ _ préféré pendant votre temps libre ? *(hobby)* 55.

56. Le jour après le lancement du nouvel iPhone 13, il voulait se p_ _ _ _ _ _ _ _ _ à la Fnac. *(haasten)* 56.

57. Le procureur a décidé la c_ _ _ _ _ _ de ce dossier de fraude. Il y a trop peu de preuves pour pouvoir condamner cet industriel. *(afsluiting)* 57.

58. Prise d'émotion, elle a versé une l_ _ _ _. *(traan)* 58.

59. Elle a décidé de mettre sa plus jolie r_ _ _ pour aller au mariage de sa soeur. *(jurk)* 59.

60. La s_ _ _ _ _ _ comprend le savoir et la vertu de quelqu'un. *(wijsheid)* 60.

Dank je wel om deze woordenschattaak te maken!

A2. Writing tasks

A2.1. 'Prize' prompts

Prompt for English:

Je hebt vannacht gedroomd dat je de **loterij had gewonnen**. Je schrijft naar een **Engelstalige vriend(in)** over je droom en wat je met dat geld allemaal zou doen.

👉 Schrijf in het **Engels** een tekst van 250-400 woorden (ongeveer **10-15 regels**).
⏰ Je krijgt daarvoor **30 minuten**.

Prompt for French:

Je hebt vannacht gedroomd dat je een prijs had gewonnen waarmee je **een jaar school mocht missen**. Je schrijft naar een **Franstalige vriend(in)** over je droom en wat je tijdens dat jaar allemaal zou gaan doen.

👉 Schrijf in het **Frans** een tekst van 250-400 woorden (ongeveer **10-15 regels**).
⏰ Je krijgt daarvoor **30 minuten**.

Translated versions:

Last night you dreamed that you had **won the lottery**. You write to an **English-speaking friend** about your dream and what you would do with the money.

👉 Write an **English** text of 250-400 words (about **10-15 lines**).
⏰ You have **30 minutes** for this task.

Last night you dreamed that you had won a prize allowing you to **miss school for a year**. You write to an **French-speaking friend** about your dream and what you would do during that year.

👉 Write a **French** text of 250-400 words (about **10-15 lines**).
⏰ You have **30 minutes** for this task.

A2.2. 'Complaint' prompts

Prompt for English:

Een bekend muziektijdschrift blijft **jouw favoriete artiest maar negeren**. Jij vindt dat die artiest nu ook eens eindelijk een plekje in het magazine verdient. Je schrijft naar een **Engelstalige vriend(in)** om je hart te luchten over de situatie.

👉 Schrijf in het **Engels** een tekst van 250-400 woorden (ongeveer **10-15 regels**).
⏰ Je krijgt daarvoor **30 minuten**.

Prompt for French:

Een tv-zender zal binnenkort **jouw favoriete programma niet meer uitzenden**. Jij vindt dat het programma absoluut wel op tv moet blijven komen. Je schrijft naar een **Franstalige vriend(in)** om je hart te luchten over de situatie.

👉 Schrijf in het **Frans** een tekst van 250-400 woorden (ongeveer **10-15 regels**).
⏰ Je krijgt daarvoor **30 minuten**.

Translated versions:

A famous music magazine keeps **ignoring your favorite artist**. You think that artist deserves to finally have a spot in the magazine. You write to an **English-speaking friend** to vent about the situation.

👉 Write an **English** text of 250-400 words (about **10-15 lines**).
⏰ You have **30 minutes** for this task.

A TV channel will soon stop broadcasting your favorite TV program. You think the program should absolutely continue to air. You write to an **French-speaking friend** to vent about the situation.

👉 Write a **French** text of 250-400 words (about **10-15 lines**).
⏰ You have **30 minutes** for this task.

A3. Questionnaires

A3.1. Questionnaire – original[48]

Met dit onderzoek willen we niet enkel zicht krijgen op je taalvaardigheid, maar ook op **welke factoren je taalvaardigheid beïnvloeden**. Met deze vragenlijst proberen we een beeld te krijgen van enkele factoren die een rol kunnen spelen.

Eerst krijg je enkele vragen over je **talenkennis**. In het tweede deel volgen enkele vragen over de **taallessen** die je op school al hebt gehad en over **jouw houding** ten opzichte van het Engels. Tot slot krijg je nog enkele vragen over de manieren waarop jij Engels tegenkomt tijdens **buitenschoolse activiteiten**.

Er zijn geen goede of slechte antwoorden. Beantwoord de vragen gewoon zo **eerlijk en volledig** mogelijk. Alvast bedankt om mee te werken!

Voor- en achternaam: ……

[A] Talenkennis

1. Welke taal/talen spreek je **thuis met je gezin**? (bv. Nederlands, Frans, Engels, Turks, Pools, Arabisch …)

……

2. Welke **andere** talen heb je al geleerd op **school**?

……

3. Heb je al eens een **taalkamp** gevolgd? Zo ja, voor welke taal?

……

4. Heb je al eens **Duolingo** gebruikt om een taal te leren? Zo ja, welke taal?

……

[48] This section shows the original version of the questionnaire for L2 English, as it was presented to the participants. For an English translation of the questionnaire, see Appendix A3.2. The questionnaire for French was nearly identical to that for English, except that occurrences of the word 'English' or 'English-language' were replaced by 'French' and 'French-language'. The questionnaire is therefore only presented in the English version.

[B] Engels taalonderwijs

1. Wanneer kreeg je **voor het eerst Engels op school**?:

□ in het eerste jaar secundair onderwijs

□ in het tweede jaar secundair onderwijs

□ op een ander moment: .....

2. Hoeveel **lesuren Engels** krijg je dit jaar in een **gemiddelde week**?

..... lesuren per week

3. Hoeveel procent van de lestijd spreekt je **leerkracht** ongeveer Engels tijdens de **Engelse les**? Zet een kruisje op de lijn.

0% ——— 100%

(nooit Engels) 25% 50% 75% (altijd Engels)

4. Volg je **CLIL**-lessen in het **Engels** (vb. geschiedenis of aardrijkskunde in het Engels)? □ ja □ nee

Zo ja, voor welke vakken?

.....

5. Waar spendeer je de **meeste tijd** aan in de **Engelse les**? Rangschik de volgende activiteiten van 1 (waar de meeste tijd naartoe gaat) naar 7 (waar de minste tijd naartoe gaat).

	Grammatica
	Lezen
	Literatuur en cultuur
	Luisteren
	Schrijven
	Spreken (individueel en interactief)
	Woordenschat

6. Hoeveel **schrijftaken** heb je vorig jaar ongeveer voor **Engels** moeten maken (in de klas en thuis)? ..... schrijftaken

7. Hoe vaak heb je vorig jaar **feedback** gekregen op schrijftaken voor **Engels**?

bij geen enkele taak	bij enkele taken	bij de helft van de taken	bij de meeste taken	bij alle taken	Ik heb geen schrijftaken moeten maken.

8. Hoeveel van die schrijftaken heb je na feedback mogen **herwerken en opnieuw mogen indienen**? ..... schrijftaken

[C] Houding tegenover het Engels

1. Ik vind **Engels**...

1.	helemaal niet leuk	niet leuk	eerder niet leuk	eerder wel leuk	wel leuk	heel leuk
2.	helemaal niet belangrijk	niet belangrijk	eerder niet belangrijk	eerder wel belangrijk	wel belangrijk	heel belangrijk
3.	helemaal niet moeilijk	niet moeilijk	eerder niet moeilijk	eerder wel moeilijk	wel moeilijk	heel moeilijk
4.	helemaal niet beangstigend	niet beangstigend	eerder niet beangstigend	eerder wel beangstigend	wel beangstigend	heel beangstigend

2. Dat Engels in Vlaanderen als **tweede vreemde taal** wordt onderwezen, vind ik...[49]

helemaal niet goed	niet goed	eerder niet goed	eerder wel goed	wel goed	heel goed

3. Wie Engels kent, moet **geen andere talen** meer kennen.

sterk oneens	oneens	eerder oneens	eerder eens	eens	sterk eens

4. Ik wil **goed zijn** in Engels.

sterk oneens	oneens	eerder oneens	eerder eens	eens	sterk eens

5. Ik werk voor mijn **lessen Engels**...

helemaal niet graag	niet graag	eerder niet graag	eerder wel graag	wel graag	heel graag

[49] In the questionnaire for L2 French, this question was phrased as follows: 'Dat Frans in Vlaanderen als eerste vreemde taal wordt onderwezen, vind ik...'.

6. **Vermijd** je om Engels te gebruiken? □ ja □ nee
Zo ja, waarom?
.....
.....

[D] Buitenschoolse activiteiten in het Engels

1. Hoelang kijk je gemiddeld **per week** naar **Engelstalige** tv-programma's, films, tekenfilmpjes, documentaires (bv. op tv, tablet, computer, gsm …)?	
Zonder ondertiteling	..... uur ..... minuten
Met ondertiteling in het **Nederlands**	..... uur ..... minuten
Met ondertiteling in het **Engels**	..... uur ..... minuten
Met ondertiteling in een **andere taal**	..... uur ..... minuten
Wat voor programma's/films kijk je het liefst/vaakst?	

2. Hoelang kijk je gemiddeld **per week** naar YouTube-filmpjes in het **Engels**?	
..... uur ..... minuten	
Wat voor filmpjes of welk kanaal bekijk je het vaakst?	

3. Hoelang hoor of lees je gemiddeld **per week Engels** op sociale media (bv. Instagram, TikTok, Facebook …)?
..... uur ..... minuten

4. Hoelang speel je gemiddeld **per week** computerspelletjes in het **Engels** (op computer, tablet, Playstation, Xbox, gsm …)?	
..... uur ..... minuten	
Welke spelletjes speel je het liefst?	

5. Hoelang luister je gemiddeld **per week** naar **Engelstalige** muziek (bv. via de radio, YouTube …)?
..... uur ..... minuten

6. Hoelang lees je gemiddeld **per week** in het **Engels** (boeken, strips, e-boeken, kranten, tijdschriften, internet …)?	
..... uur ..... minuten	
Wat lees je dan vooral?	

7. Hoelang spreek of schrijf je gemiddeld **per week** in het **Engels**…	
privé met familie, vrienden of online (bv. chatten, via sociale media, briefjes schrijven)?	..... uur ..... minuten
om zelf iets te maken (bv. verhalen schrijven, TikToks maken, muziekteksten schrijven)?	..... uur ..... minuten
Wat doe je dan zoal?	

Kom je nog op **andere manieren** met het **Engels** in contact (bv. taalkamp, internet, op reis, podcasts …)?

□ ja □ nee

Indien ja, hoe?

.....
.....

Dank je wel voor je antwoorden!

A3.2. Questionnaire – translation

With this study, we want to investigate not only your language proficiency, but also **which factors impact your language proficiency**. This questionnaire will shed light on some factors that may play a role.

First you will receive some questions about your **language knowledge**. In the second and third part, you will receive questions about **language classes** that you have taken at school and **your attitudes** towards English. Finally, there will be some questions about the ways in which you encounter English during **out-of-school activities**.

There are no right or wrong answers. Answer the questions as honestly and completely as possible. Thank you in advance for your cooperation!

First and last name: .....

[A] Language knowledge

1. Which language(s) do you speak **at home with your** family? (e.g., Dutch, French, English, Turkish, Polish, Arabic ...)

.....

2. Which **other** languages have you learned at **school**?

.....

3. Have you ever participated in a **language camp**? If yes, for which language?

.....

4. Have you ever used **Duolingo** to learn a language? If yes, which language?

.....

[B] English formal instruction

1. When did you have your **first English class in school**?:

□ in grade seven (secondary school)

□ in grade eight (secondary school)

□ in a different grade: .....

2. How many **class hours of English** do you have in a typical week this school year? ..... class hours per week

3. What percentage of class time does your teacher on average speak English during **English** class? Mark an X on the scale.



(no English) 25% 50% 75% (all English)

4. Do you take **CLIL** courses in **English** (e.g., history or geography in English)? □ ja □ nee

If yes, which courses?

.....

5. What do you spend the **most time** on in **English class**? Order the following activities from 1 (most time spent) to 7 (least time spent).

	Grammar
	Reading
	Literature and culture
	Listening
	Writing
	Speaking (individual and interactive)
	Vocabulary

6. How many **English writing tasks** did you have to complete last school year (in class and at home)? ..... writing tasks

7. How many times did you receive **feedback** on **English writing tasks**?

for no tasks	for multiple tasks	for half of the tasks	for most of the tasks	for all of the tasks	I did not have to complete

					any writing tasks.

8. How many of those writing tasks were you allowed to **revise and resubmit** after feedback? ..... writing tasks

[C] Attitudes towards Engels

1. I think **English** is…

1.	not at all fun	not fun	rather not fun	rather fun	fun	very fun
2.	not at all important	not important	rather not important	rather important	important	very important
3.	not at all difficult	not difficult	rather not difficult	rather difficult	difficult	very difficult
4.	not at all scary	not scary	rather not scary	rather scary	scary	very scary

2. The fact that English is taught as the **second foreign language** in Flanders, is…[50]

not at all good	not good	rather not good	rather good	good	very good

3. If you know English, you **don't** have to know **any other languages**.

strongly disagree	disagree	somewhat disagree	somewhat agree	agree	strongly agree

4. I want to be **good at English**.

strongly disagree	disagree	somewhat disagree	somewhat agree	agree	strongly agree

5. I like working for my **English classes**.

strongly disagree	disagree	somewhat disagree	somewhat agree	agree	strongly agree

6. Do you **avoid** using English? □ yes □ no
If yes, why?
……
……

[50] In the questionnaire for L2 French, this question was phrased as follows: 'The fact that French is taught as the first foreign language in Flanders, is …'.

[D] Out-of-school activities in English

1. How long do you spend on average **per week** watching **English-language** TV programs, movies, cartoons, documentaries (e.g., on TV, on a tablet, computer, phone …)?	
Without subtitles?	..... hour(s) ..... minutes
With subtitles in **Dutch**?	..... hour(s) ..... minutes
With subtitles in **English**?	..... hour(s) ..... minutes
With subtitles in **another language**?	..... hour(s) ..... minutes
What kind of programs/movies do you prefer watching/watch most often?	

2. How long do you spend on average **per week** watching YouTube videos in **English**?	
..... hour(s) ..... minutes	
Wat voor filmpjes of welk kanaal bekijk je het vaakst?	

3. How long do you spend on average **per week** hearing or reading **English** on social media (e.g., Instagram, Tiktok, Facebook …)?
..... hour(s) ..... minutes

4. How long do you spend on average **per week** playing computer games in **English** (on a computer, tablet, Playstation, Xbox, phone …)?	
..... hour(s) ..... minutes	
Which games do you prefer to play?	

5. How long do you spend on average **per week** listening to **English** music (e.g., on the radio, on YouTube …)?
..... hour(s) ..... minutes

6. How long do you spend on average **per week** reading in English (books, comic books, e-books, newspapers, magazines, internet …)?	
..... hour(s) ..... minutes	
What do you generally read?	

7. How long do you spend on average **per week** speaking or writing in **English**…	
to communicate with family, friends or online (e.g., chatting, on social media, writing notes or letters)?	..... hour(s) ..... minutes

to create your own content (e.g., writing stories, making Tiktoks, writing song lyrics)?	..... hour(s) ..... minutes
What do you generally do?	

Do you come into contact with **English** in any **other ways** (e.g., language camp, internet, travelling, podcasts …)?

□ yes □ no

If yes, how?

.....
.....

Thank you for your answers!

What's the word... Have you ever struggled to find the right word?

Vocabulary is at the heart of learning a second or foreign language (L2), as words are the basic building blocks required to understand and produce meaningful language. Vocabulary is particularly important to L2 writing, which requires learners to not only recall L2 words from memory, but also use them effectively in their output. Given the productive demands of L2 writing, it is essential to expand our understanding of L2 learners' productive vocabulary knowledge and vocabulary use in writing. After all, knowing a word does not necessarily mean being able to use it.

This PhD thesis guides you through a four-year research project (2021–2025) that investigated L2 learners' productive vocabulary knowledge and vocabulary use in L2 English and L2 French. It presents a large-scale, cross-sectional, cross-linguistic investigation of L2 learners in two grades of secondary school, who each completed a test battery in both English and French. The findings of three empirical studies shed light on L2 learners' productive vocabulary knowledge and vocabulary use across two grades and two L2s, and how these observed patterns are shaped by learner variables as well as methodological considerations.

