languages

Special Issue Reprint

Vocabulary Studies in L1 and L2 Development

The Interface between Theory and Practice

Edited by
Jon Clenton, Gavin Brooks and Batia Laufer

mdpi.com/journal/languages

MDPI

Vocabulary Studies in L1 and L2 Development: The Interface between Theory and Practice

Vocabulary Studies in L1 and L2 Development: The Interface between Theory and Practice

Guest Editors

Jon Clenton
Gavin Brooks
Batia Laufer

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors
Jon Clenton
The Graduate School of Humanities and Social Sciences
Hiroshima University
Hiroshima
Japan

Gavin Brooks
Faculty of Foreign Studies
Kyoto Sangyo University
Kyoto
Japan

Batia Laufer
The Department of English Language & Literature
University of Haifa
Haifa
Israel

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Languages* (ISSN 2226-471X), freely accessible at: https://www.mdpi.com/journal/languages/special_issues/Q3BDF5D230.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-4817-1 (Hbk)
ISBN 978-3-7258-4818-8 (PDF)
https://doi.org/10.3390/books978-3-7258-4818-8

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editors . . . . . **vii**

Preface . . . . . **ix**

Jon Clenton, Gavin Brooks and Batia Laufer
Vocabulary Studies in L1 and L2 Development: The Interface Between Theory and Practice
Reprinted from: *Languages* **2025**, *10*, 81, https://doi.org/10.3390/languages10040081 . . . . . **1**

Paul Nation
Re-Thinking the Principles of (Vocabulary) Learning and Their Applications
Reprinted from: *Languages* **2024**, *9*, 160, https://doi.org/10.3390/languages9050160 . . . . . **5**

Paul Meara
Thirty Years on: A Bibliometric Analysis of L2 Vocabulary Research Published in 2020
Reprinted from: *Languages* **2024**, *9*, 190, https://doi.org/10.3390/languages9060190 . . . . . **19**

Dan Hougham, Jon Clenton, Takumi Uchihara and George Higginbotham
The Impact of Lexical Bundle Length on L2 Oral Proficiency
Reprinted from: *Languages* **2024**, *9*, 232, https://doi.org/10.3390/languages9070232 . . . . . **38**

Zhiqing Li, Janis Zhiyou Li, Xiaofang Zhang and Barry Lee Reynolds
Mastery of Listening and Reading Vocabulary Levels in Relation to CEFR: Insights into Student Admissions and English as a Medium of Instruction
Reprinted from: *Languages* **2024**, *9*, 239, https://doi.org/10.3390/languages9070239 . . . . . **68**

Motoko Asano, Kensuke Hirosuna and Miho Fujieda
Exploring Lexical Bundles in the Move Structure of English Medical Research Abstracts: A Focus on Vocabulary Levels
Reprinted from: *Languages* **2024**, *9*, 281, https://doi.org/10.3390/languages9090281 . . . . . **86**

Michael Daller and Zehra Ongun
Size Matters: Vocabulary Knowledge as Advantage in Partner Selection
Reprinted from: *Languages* **2024**, *9*, 297, https://doi.org/10.3390/languages9090297 . . . . . **131**

Naheen Madarbakus-Ring and Stuart Benson
TED Talks and the Textbook: An In-Depth Lexical Analysis
Reprinted from: *Languages* **2024**, *9*, 309, https://doi.org/10.3390/languages9100309 . . . . . **145**

Roopa Leonard, Holly Joseph and Michael Daller
The Influence of the L1 on L2 Collocation Processing in Tamil-English Bilingual Children
Reprinted from: *Languages* **2024**, *9*, 319, https://doi.org/10.3390/languages9100319 . . . . . **163**

Haidee Thomson
Improving Classroom Teaching and Learning of Multi-Word Expressions for Conversational Use Through Action Research with Learner Feedback
Reprinted from: *Languages* **2024**, *9*, 336, https://doi.org/10.3390/languages9110336 . . . . . **177**

David Gregory Coulson
Retelling of Stories with Common Phrasal Expressions by High-Proficiency Learners: Implications for Learning and High-Stakes Testing
Reprinted from: *Languages* **2024**, *9*, 337, https://doi.org/10.3390/languages9110337 . . . . . **192**

Gavin Brooks, Jon Clenton and Simon Fraser
Exploring Potential Factors Affecting Reading Comprehension in EAL Learners: A Preliminary Corpus-Based Analysis
Reprinted from: *Languages* **2025**, *10*, 30, https://doi.org/10.3390/languages10020030 . . . . . **207**

Simon Fraser, Marshall Kiyoshi Higa and Walter Davies
Delivering an ESP Pedagogic Word List: Integrating Corpus Analysis, Materials Design, and Software Development
Reprinted from: *Languages* **2025**, *10*, 46, https://doi.org/10.3390/languages10030046 . . . . . **229**

About the Editors

Jon Clenton

Jon's research encompasses a diverse array of topics within applied linguistics, with a particular emphasis on vocabulary acquisition and assessment in second language (L2) contexts. His scholarly pursuits integrate theoretical frameworks and empirical methodologies to advance understanding of vocabulary development and its implications for language learning and teaching. Central to his research is the investigation of productive vocabulary knowledge and its assessment, refining methodologies for measuring productive lexical competence with a focus on isolating distinct components of vocabulary production. His work has contributed significantly to our understanding of lexical acquisition and assessment practices, exploring second language acquisition, word association studies, lexical processing, and L2 assessment. Jon's research extends to the relationship between vocabulary knowledge and language fluency, as well as lexical competence across language skills. His contributions have been published in journals such as *Bilingualism Language and Cognition*, *TESOL Quarterly*, *Language Testing*, *System*, *The International Journal of Bilingual Education and Bilingualism*, and *Language Teaching and Research*. He is a co-author of "First Language Influences on Multilingual Lexicons" (Routledge, 2020), "Vocabulary and the Four Skills" (Routledge, 2021), and "EAL Research for the Classroom, Practical and Pedagogical Implications" (Routledge, 2023), bridging research and practice in multilingual contexts. Through his comprehensive body of work, Jon has made significant contributions to vocabulary acquisition, assessment, and instruction, providing valuable insights and practical tools for language educators and researchers.

Gavin Brooks

Gavin's research focuses on corpus linguistics, vocabulary acquisition, and identifying lexical gaps in learner knowledge. Central to his work is the development of tools to support language teachers and students, such as the creation of a wordlist for EAL learners. His research focuses on using corpora and modern research methodologies to bridge the gap between theory and practice in applied linguistics and language education. Gavin is a contributing author to *International Perspectives on Teaching the Four Skills in ELT* (Palgrave Macmillan, 2018) and editor of *EAL Research for the Classroom: Practical and Pedagogical Implications* (Routledge, 2023). His work has been published in journals such as *Languages and Studies in Second Language Learning and Teaching*, contributing to advances in vocabulary learning and assessment and corpus-based research.

Batia Laufer

Batia's research focuses on different aspects of L2 vocabulary acquisition and teaching: task effectiveness, vocabulary contribution to reading and writing, lexical development, word difficulty, derivational knowledge, vocabulary testing, dictionary use, lexicography, cross linguistic influence, and attrition. She has lectured, supervised research and published widely on these topics.

Preface

This Special Issue presents a curated collection of research advancing vocabulary studies through theoretical and practical lenses. Focused on L1 and L2 development, the Special Issue unites twelve papers addressing lexical acquisition, cognitive processes, and pedagogical innovation. Motivated by the need to bridge research and classroom practice, the contributions explore cross-linguistic transfer, neurocognitive foundations of word learning, and technology-enhanced instruction.

The studies employ diverse methodologies, from corpus analyses to neuroimaging, to examine how vocabulary knowledge shapes language proficiency. Key themes include the interplay between first and second language lexicons, the role of cognitive factors in word retention, and evidence-based strategies for vocabulary assessment.

We extend gratitude to the international scholars whose work enriches this volume, offering actionable insights for educators and researchers in applied linguistics. We are especially grateful to the peer reviewers and editorial team members for ensuring scholarly rigour. We hope that this Special Issue will serve as a vital resource for academics, language teachers, and graduate students seeking to deepen their understanding of lexical development across multilingual contexts.

Jon Clenton , Gavin Brooks, and Batia Laufer
Guest Editors

languages

MDPI

Editorial

Vocabulary Studies in L1 and L2 Development: The Interface Between Theory and Practice

Jon Clenton [1,*], Gavin Brooks [2] and Batia Laufer [3]

[1] Graduate School of Humanities and Social Sciences, Hiroshima University, Hiroshima 739-8527, Japan
[2] Faculty of Foreign Studies, Kyoto Sangyo University, Kyoto 603-8047, Japan; gavinbrooks@gmail.com
[3] The Department of English Language & Literature, University of Haifa, Haifa 3498838, Israel; batialau@research.haifa.ac.il
* Correspondence: jclenton@hiroshima-u.ac.jp

The field of vocabulary studies in first language (L1) and second language (L2) development has seen remarkable growth in recent years, with researchers and practitioners alike recognizing the critical role that lexical knowledge plays in language proficiency. This Special Issue brings together innovative research that explores the intricate relationship between vocabulary acquisition theory and practical applications in both L1 and L2 contexts.

Recent studies have highlighted the importance of vocabulary knowledge in language development and its impact on various aspects of language proficiency. Zhang and Zhang (2020) conducted a meta-analysis that found a strong correlation between vocabulary knowledge and reading comprehension, showing that vocabulary accounts for a significant portion of the variance in L2 reading comprehension. These findings underscore the critical role of vocabulary in language comprehension and production, emphasizing the need for continued research and a pedagogical focus on vocabulary acquisition.

Several other recent studies have further illuminated the complexities of vocabulary acquisition and its relationship to overall language proficiency. For instance, Elgort et al. (2018) investigated the effects of contextual diversity on incidental vocabulary learning, demonstrating that exposure to words in varied contexts enhances both form and meaning acquisition. Their findings have important implications for vocabulary instruction and materials design in L2 classrooms.

In another significant contribution to the field, Pellicer-Sánchez (2020) explored the role of multimodal input in vocabulary learning, showing that the combination of textual and pictorial information can significantly enhance word learning and retention. This research highlights the potential of multimedia resources in vocabulary instruction and opens up new avenues for integrating technology into language learning environments.

Additionally, Webb et al. (2020) examined the relationship between vocabulary knowledge and listening comprehension, revealing that vocabulary size is a strong predictor of L2 listening ability. Their study emphasizes the importance of developing a robust vocabulary base to support overall language proficiency, particularly in receptive skills.

The contributions to this Special Issue build upon the central themes from this diverse array of studies and showcase the breadth and depth of current vocabulary research, spanning topics from cognitive processes in word learning to innovative pedagogical approaches. The first paper in the Special Issue by Nation provides a comprehensive framework for understanding how vocabulary "sticks" in memory, outlining eight key principles that apply to both incidental and deliberate learning. These principles—focus, accuracy, repetition, time-on-task, elaboration, analysis, motivation, and self-efficacy—offer

a clear view of what is needed for learning to occur from the viewpoint of attention. This theoretical foundation sets the stage for many of the empirical studies that follow and provides valuable insights for practitioners seeking to enhance vocabulary instruction.

This is followed up by Meara's bibliometric analysis, which offers valuable insights into the evolving landscape of L2 vocabulary acquisition research, highlighting the enduring influence of seminal scholars while also pointing to emerging areas of focus. By examining author co-citations in publications from 2020, Meara identifies Paul Nation as the most significant influence, cited in 85% of the publication set, along with other important figures such as Laufer, Hulstijn, Schmitt, and Webb. This bird's-eye view of the field helps to contextualize the more specific studies presented in subsequent articles and provides a historical perspective on the development of vocabulary research.

Several contributions examine vocabulary acquisition through the lens of multi-word units, reflecting a growing recognition that vocabulary knowledge extends beyond individual words to encompass larger linguistic chunks. Hougham et al. investigate the relationship between lexical bundles (LBs) and oral proficiency among L2 learners of varying proficiency levels at a UK university. The findings highlight the pedagogical potential of a nuanced approach to the strategic deployment of LBs, particularly bigram Mutual Information (MI) scores, to foster oral proficiency.

Asano et al. further explore the importance of lexical bundles in a different context, analyzing multi-word sequences in a one-million-word corpus of English-language medical research article abstracts. Their study reveals the importance of formulaic language in these domain-specific texts, revealing that the most frequent lexical bundles, such as "the primary endpoint was," often occurred once per text and predominantly took part in realizing a specific move within the abstract structure. The researchers found that while the bundles were comprised of relatively accessible word items, they were often technical in nature and represent an important part of how research-related discourse is conducted in this highly specific context. This study offers valuable insights into the need for learners to familiarize themselves with discipline-specific bundles to understand conventions, such as guideline adherence and statistical procedures in academic writing in the medical field.

This Special Issue also addresses the critical question of how vocabulary knowledge relates to broader language proficiency measures. Li et al. explore the correlations between aural and written vocabulary mastery and CEFR levels in an English-medium instruction (EMI) institutions. Their study of 40 undergraduate students at a Macau EMI university found positive correlations between students' CEFR levels and their mastery of both receptive aural and written vocabulary levels. These findings suggest the potential inclusion of both aural and written vocabulary levels as additional indicators for ensuring student academic success in EMI institutions, with implications for student admissions, classroom teaching, and the provision of additional English courses.

Brooks et al. examine the vocabulary demands of textbooks used in international schools, highlighting the gap between English as an additional language (EAL) learners' lexical knowledge and the vocabulary required for academic success. Their study of participants from two international schools in Japan revealed that EAL learners consistently scored lower than their first-language English (FLE) and proficient L2 (PL2) peers, with fewer than 25% of EAL learners mastering the Academic Word List before Grade 12. This analysis underscores the need for better vocabulary scaffolding and support for EAL learners in the international school context.

Several innovative approaches to vocabulary instruction and assessment are presented in this Special Issue. Madarbakus-Ring et al. evaluate the use of TED Talks in English for Academic Purposes (EAP) textbooks and offer recommendations for supplementing these materials to better support vocabulary development. Their investigation of 12 TED

Talks included in the commercial textbook Keynote 2 revealed inadequate vocabulary practice, limited academic vocabulary exposure, and a lack of item repetition for learners. The study suggests the inclusion of supplementary materials and the need for a more principled process for selecting appropriate listening activities. Their research highlights the importance of providing educators with the guidance necessary to support their learners' varying vocabulary knowledge.

Fraser et al. describe the creation and implementation of a specialized medical English word list (MEWL), along with a digital tool for personalized vocabulary learning. Their research shows how specialized word lists can be integrated into learning materials and digital tools to maximize opportunities for acquisition in the context of English for specific purposes. The MEWL, developed through needs analysis at a Japanese university's medical school, is complemented by a list of word parts and provides a framework that can be used to help students to acquire complex medical terms. The researchers also describe the development of Hi-Lex, a vocabulary learning tool that analyzes texts against word lists and allows learners to create personalized word lists, along with providing them with the support necessary to understand how these words are used in context.

The collection also ventures into interdisciplinary territory, with Daller et al. exploring the role of vocabulary size in partner selection among bilingual couples. Their novel approach shows how vocabulary research can intersect with fields such as psychology, sociology, and economics. Their study of Turkish-English bilingual couples after over 15 years of marriage found that vocabulary scores correlate highly between partners. Importantly, their results show that both L1 and L2 vocabulary knowledge play crucial roles in partner selection, suggesting that shared conceptual knowledge across languages may influence attraction and compatibility. The researchers argue that vocabulary knowledge, acquired through lengthy and costly education, serves as an asset in the "biological market" of partner selection.

While this Special Issue makes significant contributions to our understanding of vocabulary acquisition and instruction, some limitations should be noted. Most studies focus on English as the target language, and further research is needed to explore vocabulary development in other L2 contexts. Longitudinal studies examining vocabulary growth over extended periods are also underrepresented, leaving questions about long-term retention and development unanswered. The interplay between vocabulary knowledge and other language skills, such as grammar and pragmatics, also warrants further investigation.

While these are limitations of the current Special Issue, they also provide promise for the future, with several promising directions for vocabulary research emerging. There is a need for more studies that bridge the gap between L1 and L2 vocabulary acquisition, exploring how insights from one domain can inform the other. The role of technology in vocabulary learning and assessment warrants further investigation, particularly considering recent advances in artificial intelligence and natural language processing. The potential of corpus-informed vocabulary instruction and the development of more sophisticated measures of productive vocabulary knowledge are also areas ripe for exploration.

As global mobility increases, research on vocabulary development in multilingual contexts will become increasingly important. Studies that examine the interaction between multiple languages in an individual's lexicon and how this affects vocabulary learning and use could provide valuable insights for language educators and policymakers. Investigating the role of vocabulary in academic and professional success across different disciplines and cultural contexts could help to inform curriculum design and language support programs.

In conclusion, this Special Issue reflects the dynamic and multifaceted nature of contemporary vocabulary studies. By bringing together diverse perspectives and methodologies, it not only advances our theoretical understanding of lexical acquisition but also

provides valuable insights for language teachers and curriculum designers. As the field continues to evolve, the interface between theory and practice will remain crucial in shaping effective approaches to vocabulary development in both L1 and L2 contexts.

Conflicts of Interest: The authors declare no conflict of interest.

References

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. [CrossRef]

Pellicer-Sánchez, A. (2020). Learning L2 collocations incidentally from reading. *Language Teaching Research*, *24*(4), 433–456. [CrossRef]

Webb, S., Sasao, Y., & Ballance, O. (2020). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, *171*(1), 1–30. [CrossRef]

Zhang, D., & Zhang, L. J. (2020). Metacognition in TESOL: Theory and practice. *ELT Journal*, *74*(4), 429–439.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Re-Thinking the Principles of (Vocabulary) Learning and Their Applications

Paul Nation

School of Linguistics and Applied Language Studies (LALS), Victoria University of Wellington, Wellington 6140, New Zealand; paul.nation@vuw.ac.nz

Abstract: Making vocabulary stick in your memory involves dedicating attention to what needs to be learned. There are three main factors involved (focus, quantity, and quality) which can be expressed as six principles (focus, accuracy, repetition, time-on-task, elaboration, and analysis). When we include motivation in this description, then there are two more principles (motivation and self-efficacy). These principles apply to both incidental and deliberate learning, and apply to a wide range of learning focuses beyond vocabulary. These principles are well supported by research evidence. We can use the principles for re-examining teaching and learning, Technique Feature Analysis, understanding research, developing autonomy in learning, guiding curriculum design, and determining future research needs. The factors and principles provide a simple and clear view of what is needed for learning to occur from the viewpoint of attention.

Keywords: memory; repetition; vocabulary learning; attention; motivation; principles of learning; spaced retrieval; elaboration; analysis

1. Learning Vocabulary

How do we make the form and the linked meaning of a word stay in our memory? There are plenty of well-proven answers to this question, and they include advice such as focus deliberately on the word, repeat the learning at least six or seven times and preferably more, practice spaced retrieval, use a mnemonic trick such as the keyword technique, word part analysis, or considering the fit of form and meaning, write the word, and so on. This is all excellent advice that has been proven to work in particular circumstances. However, these well-proven answers put a small number of very basic principles into practice. It is useful to be aware of what these principles are because knowing them can affect how well we apply them and how widely we can apply them. The principles are surprisingly simple and involve common-sense. The framework in Figure 1 presented below suggests that if we want to learn vocabulary, we need to focus on what we want to learn (focus of attention), we need to spend time focusing on what we want to learn (quantity of attention), and we can improve learning by increasing the quality of attention to what we want to learn. Each of these three aspects of attention is on a scale, and the three aspects are very closely bound to each other. There is another aspect of a different kind that affects attention and that is motivation. Motivation affects all three of the other aspects, and in Figure 1 motivation is represented by the box around the triangle.

The three small triangles represent the cognitive requirements of attention that affect learning. The thick lines joining the triangles show that each requirement of attention is necessarily connected to the other aspects. A small triangle is used to represent each requirement to show that each requirement involves a ranked cline of levels of effectiveness, from the point to the base. Focus, quantity, and quality are strongly linked to each other because to learn something there has to be something to learn (the focus), and if we attend to it, we give it some degree of quantity of attention, and this attention has some degree of quality. Because of these strong links, there is some arbitrariness in assigning applications

to each requirement. For example, retrieval, deliberate learning, and the spacing of learning could be seen as quality features rather than quantity features, although in my discussion below, they are included in quantity. Similarly, focusing attention on a particular aspect of language use could be seen as a focus feature or as a quantity feature.

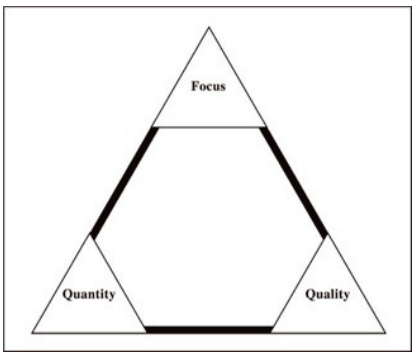

Figure 1. The aspects of attention.

As the reviewers of this paper noted, the ideas dealt with here are very much a work in progress. Publishing the paper now is in several ways premature, but at my age, you do not leave things until tomorrow. I feel sure that some of the ideas will change as I continue to try to apply them to recent research and over-looked research, and apply them from other viewpoints. However, I do feel that there is value in the reductionist and somewhat simplistic approach I have taken. Many of the ideas covered here are complex, but I think the complexity rises from the combination of principles and the circumstances of their application rather than from more complex principles I have not included. I feel the true test of the paper may be seeing if there are principles that have not been included that cannot be made from the combination of the principles I have described here.

1.1. Focus

The cline of effectiveness for focus is based on the degree to which the focus is on what really needs to be learned without distraction or error. The focus can be directed towards various aspects of form, meaning, and function (see Nation 2022, Table 2.1), and towards the knowledge components of breadth, depth, strength, and integration of vocabulary knowledge (Nation 2022, chap. 2). The focus of attention is strong if the focus is on features or skills that transfer directly to language use (relevance) and if what is focused on is clear, comprehensible, and accurate (accuracy).

1.2. Quantity

Quantity of attention involves the amount of repetition and the sustaining of attention through time-on-task. Repetition most obviously relates to the number of occurrences of vocabulary or multiword units. However, it also applies more widely to grammatical features and to skill and sub-skill development in listening, speaking, reading, and writing. The more you read, the better you become at reading through repetition of the skill and sub-skills. The cline of levels of effectiveness is quantitative, based solely on the amount of repetition and the amount of attention given. The more repetition and attention, the more likely the learning is to be retained. Repetition takes account of the gradual decay of learning, with spaced repetition re-strengthening partly decayed knowledge as well as adding to the quantity of attention given to an item.

1.3. Quality

Quality of attention involves enriching knowledge through linking present learning to existing knowledge and through increasing knowledge of different aspects of an item and different parts of aspects. As Lindstromberg (2020, p. 243) notes when discussing retrieval, "what matters is not retrieval in and of itself but rather the processing that goes

on during the search for the response and the processing that goes on after the response becomes available." The cline of levels of effectiveness is qualitative, based on the number and distinctiveness of the different aspects of what it means to know a word.

These three aspects of attention and one of motivation can be expressed as eight principles, with two for each aspect. The principles described here apply equally to incidental learning and deliberate learning. The difference between incidental and deliberate learning is simply in the degree to which each principle is applied, that is, in the position on the cline involved with each aspect. That means that deliberate learning involves stronger application of the principles, increasing their effectiveness. Incidental learning involves the same principles, but they are not so strongly applied. For example, deliberate learning of vocabulary using flash cards involves a large amount of focused attention. Incidental learning through extensive reading also involves quantity of attention but not with the same amount of attention as when using flash cards.

2. The Principles of Learning

2.1. Motivation Principles (Engagement)

For learning to occur there needs to be a willingness to focus on what needs to be learned, and to give it quantity and quality of attention.

1. Motivation: The degree of engagement with the task affects the likelihood of learning occurring.
2. Self-efficacy: Our confidence in our own skills of learning affects our success in learning.

2.2. Focus Principles (Usefulness)

Learning requires giving attention to what needs to be learned.

3. Focus: We learn what we focus on, and in addition, our learning is more useful if it closely resembles the use that we need to make of what we learn (transfer-appropriate).
4. Accuracy: Our learning is more efficient if the information we are focusing on is complete, accurate, and comprehensible. Vocabulary control, easy multiple-choice glosses that are unlikely to lead to error, and teach and test are some ways of ensuring this.

2.3. Quantity Principles (Amount)

The greater and longer the attention, the stronger the learning.

5. Repetition: The more repetition, the stronger the learning.
6. Time-on-task: Quantity of attention is increased by desirable difficulty, also called the deficient processing account. Spacing, expanding spacing, the lag effect (greater spacing rather than shorter spacing), retrieval, deliberate attention, testing (rather than re-studying), multiple-choice glosses, interleaving, production, form recall vs. meaning recall, and variation ensure a greater amount of attention and a better quality of attention resulting in better long-term retention.

2.4. Quality Principles (Connections)

Elaboration and analysis increase and strengthen connections between the item to be learned and other knowledge, making it easier to access the item. Note the overlap between quantity and quality principles, particularly where variation implies repetition. Elaboration is a commonly used term with a variety of meanings. Here, it is used with the meaning of making connections between a vocabulary item and connections outside it. Analysis examines links within the lexical item.

7. Elaboration: This includes enriching the encoding of an item through variation of different modalities (spoken, written, pictorial/visual), through variation in type of use (receptive, productive), through variation of form (meeting different family members,

or different accents), through variation in meaning and reference, through variation in grammatical use, and through embedding in larger language units. Elaboration tends to make links from the word such as to other L2 words (collocations) or to L1 words or pictures or realizations, while analysis tends to look within the word or word group to examine its parts.

8. Analysis: Relating the familiar parts to the unfamiliar whole.

3. Research Evidence

Let us now look at the research evidence for each of the principles.

1 *Motivation*: As Nation and Waring (2020) show with regard to extensive reading, motivation is influenced by a wide range of factors, and there are many kinds of motivation. Lamb (2017) provides an excellent review of how teachers can motivate learners (see Martin (2020) for a response to the article). Ushioda (2020) also takes a very practical approach to motivation. Dörnyei and Ushioda (2021) describe the difficulties in defining motivation, but stress its importance in language learning.

2 *Self-efficacy*: The idea of self-efficacy originated in the work of Bandura (1997). It involves learners' beliefs in their ability to organize and carry out a specific task to achieve certain learning. It is more than just confidence, but involves a belief in one's own autonomy in learning even though the task may be challenging (Graham 2022). There are many factors affecting the development of self-efficacy, but Usher and Pajares (2008) concluded that successfully completing challenging tasks was one of the most important. Self-efficacy and language learning strategies are closely related (Kim et al. 2015), because skill and confidence in strategy use affect whether learners will persist in their learning even though the task is challenging. This principle is a useful one to have in the list because it targets autonomy in learning as a useful goal in a language course. This principle is supported by research on autonomy, language learning strategies, and self-efficacy in foreign language learning.

3 *Focus*: The research supporting this principle includes Barcroft's (2015) research on Type of Processing Resource Allocation (TOPRA), and Morris et al.' (1977) transfer-appropriate learning.

One of the implications of the TOPRA model is that the semantic processing of words, as in extensive reading, can take attention and processing resources away from the learning of word forms. On the other hand, giving attention to word forms (structural processing) will help in the learning of word forms. A large number of studies have provided support for this idea (Barcroft 2002, 2003, 2004, 2006, 2009; Kida 2022; Kida and Barcroft 2018; Wang and Pellicer-Sánchez 2022). The idea behind TOPRA is that individuals' processing capacity is limited and if they give attention to a particular language feature, this will take attention away from other features. For example, if they pay attention to meaning in context, they take attention away from learning word forms. Elgort et al. (2018) found similar effects when using word writing to give attention to form. Wang and Pellicer-Sánchez (2022) found that bilingual subtitles were not as effective as captions for word form learning, because the L1 translation took attention away from the word form (see also Peters 2019). The implication of the focus principle for learning is that learners need to be aware of what aspect of vocabulary learning they need to focus on and make sure that they provide opportunities for that learning to occur. This means for example that when conducting extensive reading, learners note unfamiliar words for later deliberate study. Wi and Boers (2024) show the advantage of sequential processing of the same words more than once so that different aspects of knowledge can receive attention.

When examining research on the levels of processing hypothesis, Morris et al. (1977) found that the effectiveness of a particular treatment depended on the measure used to measure its effectiveness. If the measure bore little relationship to the treatment, then of course the effect of the treatment was weak. If the measure resembled the treatment or matched the goals of the treatment, then the effect was strong. The same applies to learning for language use. If the learning resembles the use the learners need to make of the

language, then the learning will result in improved language use. While proficiency in one language skill, for example reading, can support proficiency in a related skill, writing, the greatest support for the use of a skill will come from practice in that skill. This is probably the strongest criticism of the comprehensible input hypothesis. Namely, you will become much better at writing or listening if you practice writing or listening than if you do not practice writing or listening and just practice reading.

4 *Accuracy*: The research supporting this principle includes research on trial-and-error vs. informed retrieval, and research on glossing. Elgort et al. (2020) found that trial-and-error guessing from context worked better for learning word meaning than an errorless treatment where the meaning was provided before reading, but in both treatments, learners were given feedback on the correct meaning. Providing a correct answer may overcome the effects of error during trial-and-error. Boers et al. (2017) concluded that collocation activities should minimize error rates and should present complete collocations rather than parts to join together. Strong and Boers (2019) found that around 25% of learners' incorrect responses to collocation activities were reproduced in the delayed post-test. The vast majority of the textbooks that they examined provided no previous instruction to guide the learners in the exercises they included, meaning that learners had to rely on trial-and-error and later feedback. Trial-and-error may be fine if the learners are highly likely to generate the correct answer or if there is feedback that quickly deals with any errors because what we learn needs to be accurate.

Studies of glossing show that providing a meaning for a word results in better learning than guessing without feedback. Yanagisawa et al. (2020) show that being able to access a correct comprehensible meaning for a word results in more vocabulary learning than unsupported guessing from context. L1 glosses are more effective than L2 glosses, presumably because they are more comprehensible.

5 *Repetition*: There is a very large amount of research showing the positive effects of repetition on learning (Kida 2022; Laufer and Rozovski-Roitblat 2015; Pellicer-Sánchez and Schmitt 2010; Uchihara et al. 2019; Waring and Takaki 2003). The general finding is that the more repetitions there are, the more likely and stronger the learning will be. However, the relationship between repetition and learning is of medium strength rather than a very strong relationship, because there are many factors affecting vocabulary learning. Nonetheless, it is a very robust relationship. There is a persistent finding that being able to recognize or recall a word meaning requires more repetitions than developing receptive familiarity with a word form (Pellicer-Sánchez 2020; Waring and Takaki 2003). Developing collocational information and knowledge of word parts requires even more repetitions.

6 *Time-on-task*: The research supporting this principle includes research on time-on-task (Folse 2006; Hill and Laufer 2003), fluency development (Beglar et al. 2012; Chung and Nation 2006), eye-tracking (Montero Perez et al. 2015; Pellicer-Sánchez 2016; Wang and Pellicer-Sánchez 2022), desirable difficulty (Bjork and Kroll 2015), incidental vs. deliberate learning (Lindstromberg 2020; Webb 2020), deliberate attention (Laufer and Rozovski-Roitblat 2015), and the spacing of learning (Underwood 1961). Beglar et al. (2012) found that the learners who made the most reading fluency gains were learners who read the most material at the right level for them, rather than the learners who simply read the most material. This finding thus also provides support for Principle 3 which involves transfer-appropriate learning. The more fluent reading that you carry out, the more fluent you become. In an eye-tracking study using bilingual subtitles, Wang and Pellicer-Sánchez (2022) found that the amount of time spent looking at the English word form in the subtitles predicted word contributed to learning gains. Attention to the L1 translations in the bilingual subtitles took time away from giving attention to the L2 form, but assisted word meaning recall. There is evidence that spacing the retrievals increases the amount of time spent on each retrieval (Godfroid et al. 2018; Koval 2019). Laufer and Rozovski-Roitblat (2015) found, as have many other studies reviewed by them, that adding a deliberate learning element greatly increases learning. Adding a deliberate element to a task also increases the time spent on the task, particularly compared to incidental learning. Laufer

and Rozovski-Roitblat (2015), however, compared the effects of quantity and quality. They concluded that "what learners do with the word may be more important than how many times they encounter it." (p. 687), indicating that the principles of elaboration and analysis have an important role to play in learning.

7 *Elaboration*: The research supporting this principle includes research on the levels of processing theory (Craik and Lockhart 1972), varied use (Joe 1998), dual coding (D'Agostino et al. 1977; Paivio 1990), the keyword technique (Pressley 1977), and etymological elaboration (Boers et al. 2007). Craik and Lockhart suggested that the only thing that mattered in learning was what occurred in the brain at the moment of learning. If what occurred was deep and thoughtful, then the learning would last. If the processing was shallow, it would not last. This theory has come under a lot of well-deserved criticism, particularly regarding how to define deep and shallow, but the point it makes is correct—the quality of processing affects retention.

Joe (1998) examined the effects of recall on learning, finding that varied recall resulted in stronger learning. Varied recall involved the learner recalling an item in a different form or meaning from which it was initially met. This variation could involve a different morphological form, a slightly different sentence context, a slight stretching of the meaning, and so on. Joe found it necessary to develop a scale of variability to classify the variations produced.

In typical receptive language use, meetings with the same word typically occur in different contexts. The occurrence of different family members from the same family, such as *involving* and *involvement*, is likely to result in stronger learning than exactly the same family member occurring again, underlining the importance of including attention to word families and word part analysis in a language course.

Paivio (1990) proposed that a critical factor affecting learning was the coding of information in two ways—a language-based (verbal) representation and non-verbal imagery. The two kinds of representation are separate systems, but they are functionally interconnected. The connections between the two systems improve the quality of processing and thus help memory.

The keyword technique is a memory trick for learning vocabulary that uses linguistic and visual information. It is the most researched mnemonic device for learning vocabulary with well over sixty research studies focusing on it. The research is reviewed in Nation (2022) and Boers (2021). The technique involves finding a first language keyword that sounds like the beginning or all of a foreign language word that needs to be learned. For example, for learning the English word *cow*, the L1 Thai word *khaw* meaning "rice" could be used. The next step is to create a visual image involving the interaction of the meanings of the two words, for example a cow eating rice, or a cow sitting on a pile of rice. The keyword technique thus involves a form link (the sounds of the two words) and a meaning imagery link. The keyword technique typically produces results around 20%-25% better than rote learning. Its effectiveness depends on it being used properly (Boers 2021, pp. 176–81) in line with the principles of learning described here.

Etymological elaboration applies to the learning of multiword units and figurative expressions (Boers et al. 2007), and there is a large amount of research showing the effectiveness of learning about the origin of expressions such as *the ball is in your court* as a way of helping them stick in memory. Such elaboration can apply to single words and multiword units with the learning of a word such as *trust* being related to *true*, *tree,* and *trunk*. Boers and colleagues' research has also shown the effect of phonological repetition such as alliteration, rhyme, and euphony in helping retain the form of phrases (Boers 2020). Similarly, when mapping the form–meaning link, encouraging the learners to consider the appropriateness of the link (does the form of this word fit its meaning?) helps in learning the link (Deconinck et al. 2017), as a result of cross-linguistic associations, or what the authors call "sound symbolic resonance".

It is important to note that the principle of elaboration works in conjunction with the other principles. What is elaborated needs to relate to the focus of learning, needs to be accurate, and needs quantity of attention.

8 *Analysis*: The research supporting this principle includes word part analysis (Wei 2015) and longitudinal and multi-aspect studies of vocabulary depth (Pigada and Schmitt 2006). There are systems that lie behind the various aspects of knowing a word. These systems include sequences of sounds, spelling rules, word parts, form–meaning connections, the shared underlying meaning of various uses of a word, the patterning of multiword units, and grammatical patterns. Seeing how particular items relate to these systems helps learning by relating familiar knowledge to new knowledge.

This reconceptualization of what we already know has several practical applications, namely for teaching and learning, Technique Feature Analysis, understanding, explaining and predicting research findings, developing learner autonomy, and directing future research. The following part of the paper is an attempt to check out whether this re-thinking has any noticeable effect on these applications or if it is just shuffling ideas around.

Application 1: Applying the principles to teaching and learning

Table 1 relates the eight principles to their applications in learning. The guidelines such as using retrieval, spaced repetition, visualization, and using mnemonics that have typically been used to enhance learning still apply. The difference is that now it is easier to see why these guidelines are relevant, and how they should be applied. Having knowledge of why the guidelines work can ensure that the guidelines are used well. For example, when conducting flash card learning, knowing that what you focus on is what you are most likely to learn can help a learner decide if they should say the words aloud, write them, or even close their eyes and visualize the word form before retrieving the meaning. It should also make them aware that when the words on the cards are getting easier after several repetitions, it is still worth dwelling on them rather than rushing through. The applications described in column 2 are given as very direct statements for the sake of clarity. However, they need to be applied with a little more nuance. For example, for Principle 4 *Use dictionaries and glossaries*, the glossaries and dictionaries should be comprehensible, either using L1 explanations or explanations in a controlled vocabulary. Similarly, *Don't use trial-and-error* can be qualified by saying "unless the chance of error is very small". Typically, the things to be careful about involve the application of principles which are in the list of eight.

Table 1. Principles and their applications for vocabulary learning.

Principle	Application in Techniques
1 *Motivation*	Learn obviously useful words Set challenging but achievable goals Use puzzle-like and test-like tasks Record successful learning using graphs
2 *Self-efficacy*	Use vocabulary control in listening and reading materials Do pre-study training Work with small numbers of words
3 *Focus*	Understand what is involved in understanding a word Have a clear vocabulary learning goal Supplement incidental learning with deliberate learning
4 *Accuracy*	Do not use trial-and-error, use study and test Use comprehensible dictionaries or glossaries Use L1 meanings

Table 1. *Cont.*

Principle	Application in Techniques
5 *Repetition*	Use peer testing Use related tasks that recycle the same content Encourage large quantities of graded reading and listening Encourage opportunities for negotiation of meaning Revisit the same material several times
6 *Time-on-task*	Use spaced retrieval Use self-testing Use larger groups of cards as skill at learning increases
7 *Elaboration*	Find members of the same word family Look for the core meaning of various senses Learn words with their known collocations Conduct plenty of extensive reading to see examples of use Focus on both written and spoken forms Say or write the words, preferably using retrieval Visualise examples of use Use the words Use pictures of the meaning as well as translations Practice receptive and productive learning, that is looking at the word and recalling its meaning, and looking at the meaning and recalling the word form Consider the fit of form and meaning Look for cross-linguistic associations
8 *Analysis*	Learn to use the keyword technique Learn word parts and learn to recognize them in words Do multiword unit analysis Consider the origins of figurative expressions Use analogy or patterning of form, meaning, and use. For example, look at spelling regularities and irregularities, look for a shared meaning among various uses, look for similarities among collocates

As Bjork et al. (2013) note, learners and teachers are often not good judges in deciding on the success of their learning. This unsuccessful estimation typically occurs because a short-term gain does not necessarily result in a long-term gain, or because initial difficulty may actually help long-term retention. Relying on the application of principles makes the use and adaptation of learning activities more effective than simply relying on intuition.

The relationship of a task to previous learning strongly affects the way the task contributes to learning. For example, verbatim repetition is likely to contribute less than varied repetition. Note also that analysis and elaboration enrich knowledge of the item and also increase the amount of time focused on an item.

Application 2: Revised Technique Feature Analysis

Technique Feature Analysis (Nation and Webb 2011) is a more elaborated form of the Involvement Load Hypothesis (Laufer and Hulstijn 2001). It was created because the Involvement Load Hypothesis does not include several conditions that have been proven to help vocabulary learning, such as varied use, repetition, spacing, and analysis. Technique Feature Analysis tries to predict the effectiveness of a vocabulary learning activity by analysing how well the activity involves established principles of learning. Reconceptualizing the principles of learning thus necessitates reconceptualizing Technique Feature Analysis. The result of this reconceptualization is that there are now 16 questions instead of 18. The deleted features are "Awareness of new vocabulary learning" (because it is covered under "Focus attention on what we need to learn"), and "Negotiation" (because it is covered under "Varied meetings and use"). The headings in Figure 2 match the four parts of Figure 1—motivation, focus, quantity, and quality.

Motivation
Are the learners likely to be successful in completing the activity?
Does the activity motivate *attention or engagement?*
Do the learners select the words?
Focus
Does the activity focus attention on *what we need to learn*?
Is the learning *transfer-appropriate*?
Is the information *accurate and comprehensible*?
Quantity
Does the activity involve retrieval of the meaning *or substantial attention to the meaning*?
Does the activity involve retrieval of the word form or substantial attention to the word form?
Is it recall rather than recognition?
Are there multiple meetings or retrievals of each word?
Is there spacing *of at least several minutes* between meetings or retrievals?
Quality
Does the activity involve varied meetings or use?
Is there a substantial change from previous meetings?
Does the activity involve analysis or analogy?
Does the activity involve instantiation or imaging?
Does the activity avoid interference?
Total score/16

Figure 2. Technique feature analysis. *Italics* = new additions.

The following comments refer to the questions in the table. *Does the activity motivate attention or engagement* can include interest in a story. *Do the learners select the words* could include deciding which words to look up. *Does the activity involve varied meetings or use* can include variation from previous meetings or within the activity itself as with varied word forms, collocations, contexts. *Is there substantial change from previous meetings* probably needs a two- or three-point scaled list, as in Joe (1998). One or two other criteria might benefit from being a scale rather than being just yes/no. Multiple meetings could be scaled on the number of repetitions. Length of spacing could be scaled on length of time. Does the activity motivate attention could also be scaled.

Here is an example of applying the questions using the keyword technique. The keyword technique would be scored two out of three for motivation if the learners are well trained in its use (the activity is not self-motivating). It would be scored two out of three for focus (It is not transfer-appropriate). It would be scored two out of five for Quantity (It is recognition not recall. It does not involve repetition or spacing). It would be scored three out of five for quality (It does not involve variation or substantial change). It would have a total of 9 out of 16, a rather harsh score perhaps.

Applying the re-thinking to Technique Feature Analysis results in several important deletions, additions, and adjustments. It needs further adjustments to preserve its yes/no nature.

Application 3: Understanding and predicting research findings

The motivation for the writing of this paper came from my rather arrogant belief that if you showed me the treatment and measurement sections of a piece of vocabulary research, I could successfully predict what the results would be. This interest partly arose from reading papers where time-on-task was not controlled for and studies comparing incidental and deliberate learning. I was going to give the paper the tongue-in-cheek title of *Do we need more research?*, working on the idea that if we could predict results from already well-proven principles, then we should be researching areas where we do not know what the results will be. I started testing myself out on papers where I knew the result and then

on experiments and meta-analyses where I deliberately avoided looking at the results. I found my predictions were largely correct.

Such prediction is a form of Technique Feature Analysis, where a technique is given a score on the basis of meeting several learning criteria. When analysing an experiment, the analysis is made quite a bit easier by having a comparison with a control group to highlight the critical part of the treatment. In most research, we tend to work the other way. We look at the results, and then in the discussion section we explain the results by relating them to principles and previous research. Some pieces of research make predictions before gathering the data. I found I was quite good at predicting the result, and that made me look again at what principles of learning I was drawing on to make the prediction. Being able to make the predictions also made me consider whether we were doing too much research in areas where there were few new things to learn, and not enough research involving brave questions where the answers are truly unknown. Such brave research, where we do not know the answers, includes the development of implicit vocabulary knowledge, the effect of fluency development activities on collocational knowledge, the growth of L1 vocabulary knowledge during tertiary study and in employment, the growth of L2 word family knowledge, and the effect of extensive reading on knowledge of grammar.

Understanding the principles also helps us make sense of the research we read. If when reading we can compare what we read with what we already know about learning, such as the principles described here, then our reading will be more informed and more critical.

Application 4: Learner autonomy

Learning how to learn is neglected in many language courses, and certainly in most published language courses. However, learning a language is such a large undertaking that it necessarily requires learners to take control of their own learning because coursebooks and teachers cannot possibly cover all that needs to be learned, and much of the learning involved is beyond the influence of direct teaching and study. Learners need to know how to learn.

The framework and principles described in this paper are not difficult to understand, and they require only a small amount of learning. When training learners in how to learn, there is value in having a simple clear memorable framework that they can follow. Figure 1 fulfils these requirements. Learning how to learn includes understanding the principles behind learning.

A useful way to learn the principles would be to analyse applications, such as those in Table 1, using the principles, and to improve learning procedures using the principles. The principles apply well beyond vocabulary learning.

Application 5: The principles of course design

How does re-thinking the principles of learning affect the wider set of principles of course design? Macalister and Nation (2020) present 20 principles of course design organised under the three headings of content and sequencing, format and presentation, and monitoring and assessment.

The learning principles not obviously covered in the larger set of principles of course design are the focus principles of focus and accuracy. It could be argued that the format and presentation principles of the four strands and the related principles of comprehensible input, fluency, output, and deliberate learning to some degree cover focus in that they deal with transfer-appropriate learning through input, output, and fluency development, and that deliberate learning involves deliberate focus, but this obscures the focus principle which emphasizes that we need to focus on what we need to learn. The accuracy principle is completely missing from the original list of curriculum design principles.

An improved list of curriculum design principles would look like this (Table 2), with omissions to keep the list to 20 principles.

Table 2. The adjusted list of principles of curriculum design.

Principles of Course Design	Principles Removed or Reformed
Content and sequencing Cost/benefit *Focus* *Self-efficacy* *Accuracy* *Repetition* Language system Keep moving forward Learning burden Interference	Frequency Strategies and autonomy Spaced repeated retrieval
Format and presentation *Motivation* Four strands Comprehensible input Fluency Output Deliberate learning *Time-on-task* *Elaboration* *Analysis*	Depth of processing
Monitoring and assessment Ongoing needs and environment analysis Feedback	

The eight principles of learning are in italics.

The principles of teachability, integrative motivation, and learning preferences have been removed from the list of curriculum design principles, partly because research has not strongly supported their inclusion. A further content principle to consider including is part of a revision of the four strands principle (Nation 2017), namely integrated content, though this could be considered as part of the four strands principle. Rethinking the principles of learning has a substantial effect on around one-third of the twenty principles of curriculum design with three deletions, three additions, and four re-namings and reformings.

Application 6: Future research

While we know a lot about quantity of attention, we know less about quality of attention and what it contributes beyond quantity. How do analysis and elaboration enhance learning? In what ways does the research on word association have application to elaboration? Does long-term storage result in more secure storage without further quantity or quality of attention needed? Do the principles apply to grammar? Quality of attention as described in this paper seems to have little relevance to the subconscious learning of grammar, except perhaps for grammatical features that have parallels in the L1. It may also have little relevance to the advanced learning of vocabulary when core meanings which are different from L1 translations are established. We know little about the development or word meaning in an L2. It is likely that the way we develop concepts in our L1 will be similar to the way advanced learners develop concepts in their L2.

4. Problems

According to this framework, if the quantity of attention is truly the same between massed learning and spaced learning, then the learning will be the same, both for short-term and long-term retention. I feel this is probably not right. Research on topics such as the effect of sleep on learning suggests that qualitative changes can occur between spaced repetitions (Dumay and Gaskell 2007), particularly in the integration of new knowledge into existing knowledge.

Using Technique Feature Analysis still requires a lot of judgement to be made in answering the questions. Considerable training may be needed to encourage raters to agree on analyses.

5. Reflections

This paper has not focused on new research but has looked at re-thinking what we already know. This re-thinking has resulted in some notable changes to Technique Feature Analysis and the principles of curriculum design, and an easily remembered mnemonic diagram for learner training. It has not resulted in much change to the way we view and explain teaching and learning activities because re-thinking the principles does not change the existing already useful and well-proven guidelines for learning. The re-thinking does affect the way we interpret and explain research findings, because research often aims to provide new guidelines. The simple framework makes explaining and predicting findings clearer because there is now a smaller number of factors to consider. Overall, I consider the re-thinking has enough positive practical outcomes to have made it worthwhile.

This re-thinking of the principles of learning has resulted in a book on the twenty most effective language teaching techniques (Nation 2024). This book involves an evaluation of very commonly used language teaching techniques and involves considering how well they apply the principles of learning and how they should be used to make sure that they make the best application of the principles. As this article suggests, there is a small number of principles of language learning that can be applied to a large number of teaching and learning procedures, and to analysing, applying, and researching these procedures.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The author declares no conflict of interest.

References

Bandura, Albert. 1997. *Self-Efficacy: The Exercise of Control*. Dallas: Freeman.

Barcroft, Joe. 2002. Semantic and structural elaboration in L2 lexical acquisition. *Language Learning* 52: 323–63. [CrossRef]

Barcroft, Joe. 2003. Effects of questions about word meaning during L2 Spanish lexical learning. *Modern Language Journal* 87: 546–61. [CrossRef]

Barcroft, Joe. 2004. Effects of sentence writing in second language lexical acquisition. *Second Language Research* 20: 303–34. [CrossRef]

Barcroft, Joe. 2006. Can writing a word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research* 22: 487–97. [CrossRef]

Barcroft, Joe. 2009. Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly* 43: 79–103. [CrossRef]

Barcroft, Joe. 2015. *Lexical Input Processing and Vocabulary Learning*. Amsterdam: John Benjamins.

Beglar, David, Alan Hunt, and Yuriko Kite. 2012. The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning* 62: 665–703. [CrossRef]

Bjork, Robert A., and Judith F. Kroll. 2015. Desirable difficulties in vocabulary learning. *The American Journal of Psychology* 128: 241–52. [CrossRef] [PubMed]

Bjork, Robert A., John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology* 64: 417–44. [CrossRef] [PubMed]

Boers, Frank. 2020. Factors affecting the learning of multiword items. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. New York: Routledge, pp. 143–57.

Boers, Frank. 2021. *Evaluating Second Language Vocabulary and Grammar Instruction*. London: Routledge.

Boers, Frank, June Eyckmans, and Hélène Stengers. 2007. Presenting figurative idioms with a touch of etymologyore than mere mnemonics? *Language Teaching Research* 11: 43–62. [CrossRef]

Boers, Frank, Tu Cam Thi Dang, and Brian Strong. 2017. Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead and Webb 2014. *Language Teaching Research* 21: 362–80. [CrossRef]

Chung, Mihwa, and I. S. Paul Nation. 2006. The effect of a speed reading course. *English Teaching* 61: 181–204.

Craik, Fergus I. M., and Robert S. Lockhart. 1972. Levels of processing framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11: 671–84. [CrossRef]

D'Agostino, Paul R., Brian J. O'Neill, and Allan Paivio. 1977. Memory for pictures and words as a function of level of processing: Depth or dual coding? *Memory and Cognition* 5: 252–56. [CrossRef] [PubMed]
Deconinck, Julie, Frank Boers, and June Eyckmans. 2017. 'Does the form of this word fit its meaning?' The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research* 21: 31–53. [CrossRef]
Dörnyei, Zoltán, and Ema Ushioda. 2021. *Teaching and Researching Motivation*, 3rd ed. London: Routledge.
Dumay, Nicolas, and M. Gareth Gaskell. 2007. Sleep-associated changes in the mental representation of spoken words. *Psychological Science* 18: 35–39. [CrossRef] [PubMed]
Elgort, Irina, Natalia Beliaeva, and Frank Boers. 2020. Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and non-declarative knowledge. *Studies in Second Language Acquisition* 42: 7–32. [CrossRef]
Elgort, Irina, Sarah Candry, Thomas J. Boutorwick, June Eyckmans, and Marc Brysbaert. 2018. Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics* 39: 464–87. [CrossRef]
Folse, Keith S. 2006. The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly* 40: 273–93. [CrossRef]
Godfroid, Aline, Jieun Ahn, Ina Choi, Laura Ballard, Yaqiong Cui, Suzanne Johnston, Shinhye Lee, Abdhi Sarkar, and Hyung-Jo Yoon. 2018. Incidental vocabulary learning in a natural reading contextn eye-tracking study. *Bilingualism: Language and Cognition* 21: 563–84. [CrossRef]
Graham, Suzanne. 2022. Self-efficacy and language learning—What it is and what it isn't. *The Language Learning Journal* 50: 186–207. [CrossRef]
Hill, Monica, and Batia Laufer. 2003. Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL* 41: 87–106. [CrossRef]
Joe, Angela. 1998. What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics* 19: 357–77. [CrossRef]
Kida, Shusaku. 2022. Secondary task type, exposure frequency and their combined effects on second language vocabulary learning through reading. *Second Language Research* 38: 213–32. [CrossRef]
Kida, Shusaku, and Joe Barcroft. 2018. Semantic and structural tasks for the mapping component of L2 vocabulary learning: Testing the TOPRA model from a new angle. *Studies in Second Language Acquisition* 40: 477–502. [CrossRef]
Kim, Do-Hong, Chuang Wang, Hyun Seon Ahn, and Mimi Bong. 2015. English language learners' self-efficacy profiles and relationship with self-regulated learning strategies. *Learning and Individual Differences* 38: 136–42. Available online: https://bit.ly/2rrvjYO (accessed on 28 January 2024). [CrossRef]
Koval, Natalie G. 2019. Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics* 40: 1103–39. [CrossRef]
Lamb, Martin. 2017. The motivational dimension of language teaching. *Language Teaching* 50: 301–46. [CrossRef]
Laufer, Batia, and Bella Rozovski-Roitblat. 2015. Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research* 19: 687–711. [CrossRef]
Laufer, Batia, and Jan Hulstijn. 2001. Incidental vocabulary acquisition in a second languagehe construct of task-induced involvement. *Applied Linguistics* 22: 1–26. [CrossRef]
Lindstromberg, Seth. 2020. Intentional L2 vocabulary learning. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. New York: Routledge, pp. 240–54.
Macalister, John, and I. S. Paul Nation. 2020. *Language Curriculum Design*, 2nd ed. New York: Routledge.
Martin, Christopher. 2020. Response to 'The motivational dimension of language teaching' (Lamb, 2017). *Language Teaching* 53: 233–36. [CrossRef]
Montero Perez, Maribel, Elke Peters, and Piet Desmet. 2015. Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal* 99: 308–28. [CrossRef]
Morris, C. Donald, John D. Bransford, and Jeffery J. Franks. 1977. Levels of processing vs. transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* 16: 519–33. [CrossRef]
Nation, I. S. Paul. 2022. *Learning Vocabulary in Another Language*, 3rd ed. Cambridge: Cambridge University Press.
Nation, I. S. Paul. 2024. *The Twenty Most Effective Language Teaching Techniques*. London: Routledge.
Nation, I. S. Paul, and Rob Waring. 2020. *Teaching Extensive Reading in Another Language*. London: Routledge.
Nation, I. S. Paul, and Stuart Alexander Webb. 2011. *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.
Nation, Paul. 2017. How vocabulary is learned. *Indonesian Journal of English Language Teaching* 12: 1–14. [CrossRef]
Paivio, Allan. 1990. *Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press.
Pellicer-Sánchez, Ana. 2016. Incidental L2 vocabulary acquisition *from* and *while* reading. *Studies in Second Language Acquisition* 38: 97–130. [CrossRef]
Pellicer-Sánchez, Ana. 2020. Expanding English vocabulary knowledge through reading: Insights from eye-tracking studies. *RELC Journal* 51: 134–46. [CrossRef]
Pellicer-Sánchez, Ana, and Norbert Schmitt. 2010. Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language* 22: 31–55.
Peters, Elke. 2019. The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly* 53: 1008–32. [CrossRef]

Pigada, Maria, and Norbert Schmitt. 2006. Vocabulary acquisition from extensive reading case study. *Reading in a Foreign Language* 18: 1–28.

Pressley, Michael. 1977. Children's use of the keyword method to learn simple Spanish vocabulary words. *Journal of Educational Psychology* 69: 465–72. [CrossRef]

Strong, Brian, and Frank Boers. 2019. The error in trial and error: Exercises on phrasal verbs. *TESOL Quarterly* 53: 289–319. [CrossRef]

Uchihara, Takumi, Stuart Webb, and Akifumi Yanagisawa. 2019. The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning* 69: 559–99. [CrossRef]

Underwood, Benton J. 1961. Ten years of massed practice and distributed practice. *Psychological Review* 68: 229–47. [CrossRef]

Usher, Ellen L., and Frank Pajares. 2008. Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research* 78: 751–96. Available online: http://www.jstor.org/stable/40071145 (accessed on 28 January 2024). [CrossRef]

Ushioda, Ema. 2020. *Language Learning Motivation*. Oxford: Oxford University Press.

Wang, Andi, and Ana Pellicer-Sánchez. 2022. Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning* 72: 765–805. [CrossRef]

Waring, Rob, and Misako Takaki. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language* 15: 130–63.

Webb, Stuart. 2020. Incidental vocabulary learning. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. New York: Routledge, pp. 225–39.

Wei, Zheng. 2015. Does teaching mnemonics for vocabulary learning make a difference? Putting the keyword method and the word part technique to the test. *Language Teaching Research* 19: 43–69. [CrossRef]

Wi, Injung, and Frank Boers. 2024. Sequential use of L1 and L2 captions: Exploring the benefits for vocabulary acquisition. *TESOL Quarterly* 58: 511–21. [CrossRef]

Yanagisawa, Akifumi, Stuart Webb, and Takumi Uchihara. 2020. How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *SSLA* 42: 411–38.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Thirty Years on: A Bibliometric Analysis of L2 Vocabulary Research Published in 2020

Paul Meara [1,2]

[1] Applied Linguistics, Swansea University, Singleton Park, Swansea SA2 8PP, UK; p.m.meara@gmail.com
[2] Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK

Abstract: This paper presents an author co-citation analysis of the research on L2 vocabulary acquisition that was published in the 2020 calendar year. The most significant influence at this time is Paul Nation—cited in 85% of the publication set—but a number of other important influences can also be identified, notably, Laufer, Hulstijn, Schmitt and Webb. This paper draws some comparisons with data from 1990, and speculates on how "research fronts" might be identified in an author co-citation data set.

Keywords: author co-citation; bibliometrics; 2020; L2 vocabulary research

1. Introduction

Over the last few years, I have published a number of papers that sketched the development of L2 vocabulary research. These analyses have included a series of maps that illustrate the co-citation patterns linking the authors who are cited in the research outputs. So far, I have produced a set of maps covering the period 1982–1991. Summaries of this work, and maps covering each of the 10 years in the period under consideration, can be found on my web-site: https://www.lognostics.co.uk/maps (accessed on 1 May 2024).

A typical map of this sort is shown in Figure 1, based on the co-citations found in the 78 papers that make up the 1990 data set.

The map consists of a set of labelled nodes and a set of edges. The nodes identify the 65 most cited authors in the 1990 data set. The edges show the pattern of co-citation among these authors. Thick edges show a high level of co-citation, thinner edges indicate fewer co-citations. Each author is connected with the author they are most frequently co-cited with. So, for example, the strongest co-citation links in this map are Gairns and Redman (cluster III), co-cited seven times in this data set, and Carter and McCarthy, co-cited six times in the data set (cluster I). A factor analysis of these data identifies ten clusters in the map, which we can interpret as "invisible colleges" (cf. Price 1965; Crane 1972)—groups of researchers who were influencing particular themes in the L2 vocabulary research activity that was published in 1990. Cluster II, for example, is a group of psycholinguists whose work is often cited in the context of bilingual lexical performance; Cluster V is a group of L1 reading researchers whose work is often cited in the context of L2 reading research; Cluster IX is a set of word lists that are frequently cited in the context of L2 readability schemes. (See (Meara 2022, 2023) for a more detailed discussion of this map).

Readers who are familiar with the early research literature on vocabulary acquisition will easily recognise that the mapping captures the main trends in the research at the time. They will also feel that the map probably looks very different from the sort of map that we might expect to emerge from an analysis of more recent data. Only three of the significant influences in the current research (Nation, Laufer, Meara) play a significant role in this map, and this raises some interesting questions about the way L2 vocabulary research has developed over the last thirty years. What new research clusters have emerged since 1990? At what point in the timeline do other significant figures begin to appear? How does their

research cluster with the work of earlier researchers? What new research themes are they associated with? What impact do these new themes have on the field as a whole?

A number of colleagues have pointed out that these early maps are mainly of historical interest, however, and some (notably Norbert Schmitt) have argued that it would be much more interesting to apply the same methodology to the more recent research output. This paper is an attempt to address this interest. It provides an analysis of the relevant research published in 2020. To cut a long story very short, the 2020 research output map does indeed look very different from the research that was published in the 1990, and it is interesting to reflect on why these differences appear in the data, and how we should understand them.

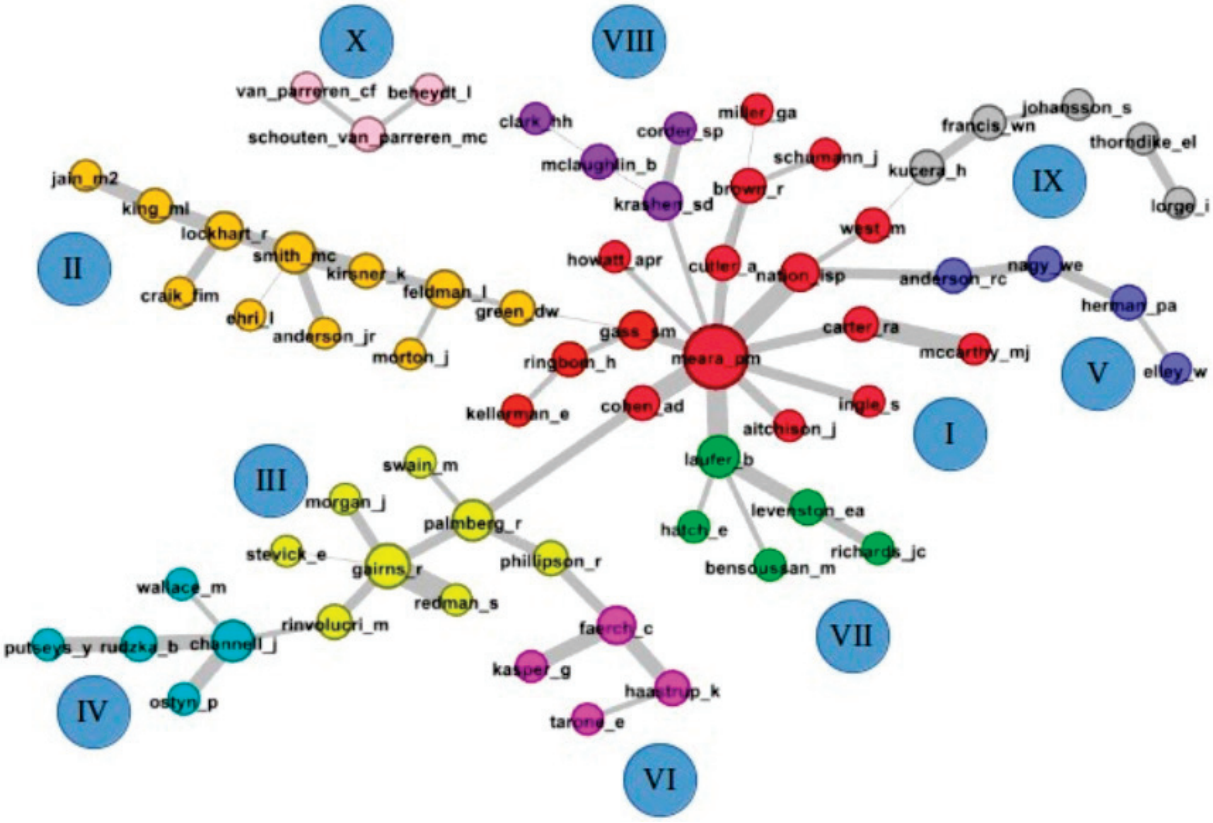

Figure 1. The 1990 co-citation data set shown as a spanning tree. Each node represents an author. Edges show the strongest co-citation connections between authors; e.g., in this tree, Carter and McCarthy are strongly co-cited, while Stevick and Gairns are co-cited only a few times.

2. The Background to the 2020 Research

Before we proceed to the 2020 data, it is worthwhile providing some contextual background that will place this research in perspective.

Figure 2 shows the number of L2 vocabulary-related research outputs identified in the Vocabulary Acquisition Research Group Archive (VARGA: Meara n.d.) during the period 1980–2023. The figure shows that relatively few papers appeared in the early years of this time span. It also shows that the number of outputs increases steadily, reaching 200 annual publications by 1993, exceeding 300 annual publications in 2010, and breaking the 400 publications barrier two years later in 2012. Looking at the raw figures, it is tempting to suggest that the number of research outputs is in decline following a peak in 2018. However, I think this interpretation of the data is incorrect. Not all publications are logged immediately after their first appearance, and it often takes several years for publications that are published in obscure journals to emerge from hiding. At the time of writing (December 2023), the VARGA database identifies a total of 318 relevant publications published in 2020 (240 journal articles, 67 book chapters, seven books/monographs, three theses, and one computer program). The VARGA database does not systematically monitor

theses, so the three works of this type that have been recorded for 2020 probably under-represent this type of output by a significant factor. Based on the number of publications appearing in the period 2015–2019, my estimate is that the figure of 318 publications appearing in 2020 represents approximately 75% of the number of publications we might expect to find, indicating that approximately a hundred publications remain to be found. This means that the analysis that follows should be considered as an interim analysis that identifies the main trends in 2020, but it remains liable to revision if some radically new sources reveal themselves at a later date.

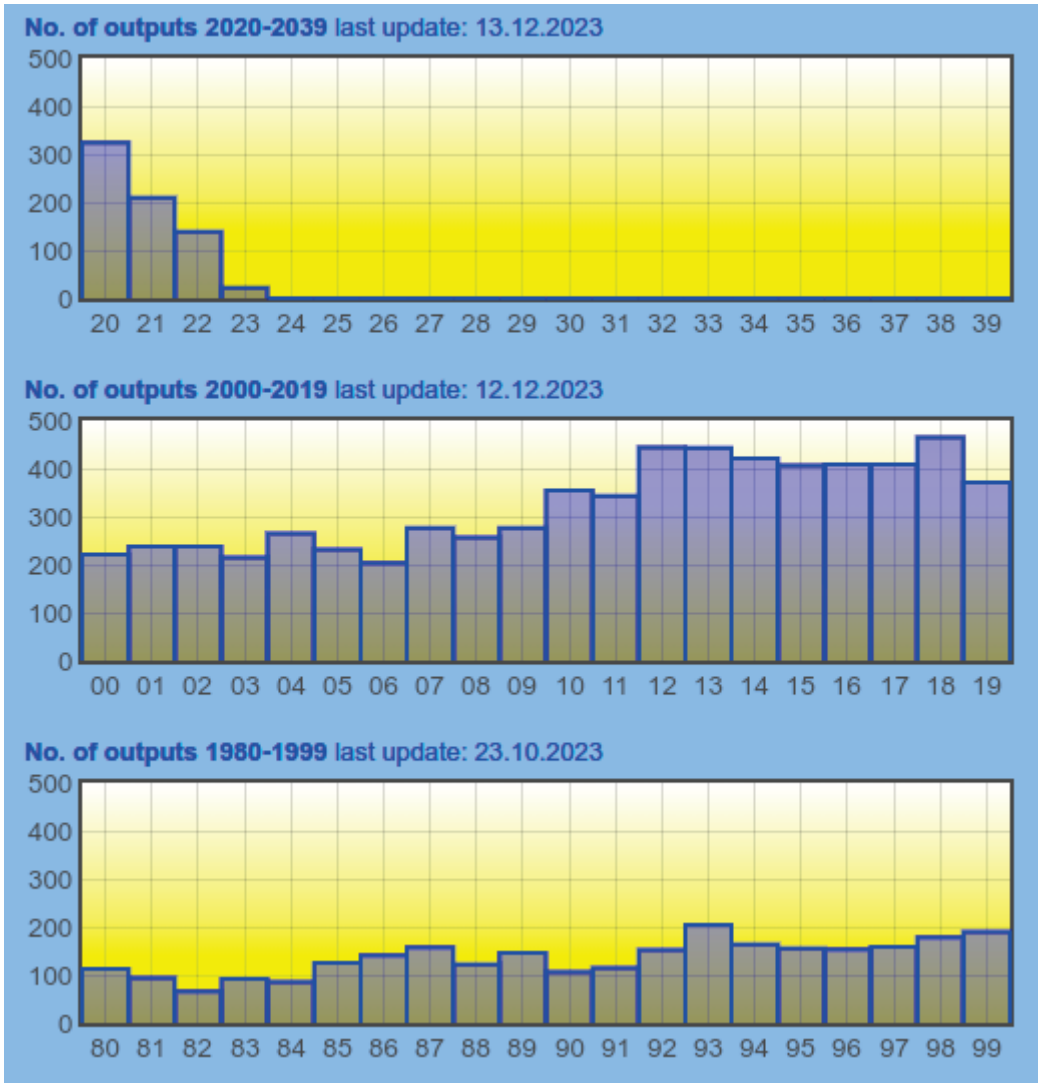

Figure 2. The number of L2 vocabulary-related research papers published in the years 1980–2023.

The interim data set is fairly typical of the research that has appeared in recent years. The vast majority of the sources identified appear as journal articles or chapters in books—307 works in total. In the interest of space, I have not listed all the 2020 sources in this paper, but curious readers can find a list of the 2020 publications and their type in the VARGA data base. https://www.lognostics.co.uk/varga/ (accessed on 1 May 2024).

A superficial analysis of the 2020 publications reveals that 493 unique authors contributed to the 2020 outputs. The vast majority of these authors contributed to only a single publication, but a handful of authors contributed to two or more publications (see Table 1). The outstandingly prolific author in 2020 is Webb, who contributed to 10 outputs. A total of 4 authors contributed to four outputs (Dang, Kyle, Peters, and Zhang) and 14 authors contributed to three outputs (Coxhead, Laufer, Lee, McLean, Miralpeix,

Pellicer-Sanchez, Reynolds, Rodgers, Stoeckel, Teng, Tremblay, Uchihara, and Yanigasawa). A further 35 authors each contributed to two papers in 2020.

Table 1. The number of authors contributing to N outputs in 2020.

Outputs	10	9	8	7	6	5	4	3	2	1
Authors 2020	1						4	13	35	439
Lotka estimate	4	5	7	9	12	18	27	49	109	

It is worth noting here that the number of authors identified in the 2020 data set is considerably larger than the number of authors appearing in the 1990 data described earlier. Only 94 authors were identified in the 1990 data set, and only 8 of those authors contributed to more than one paper. The most prolific authors in 1990 were Laufer and Meara, who both contributed four papers to the data set. In contrast, five authors published four or more studies in 2020. These figures underline the massive growth in output that has taken place between 1990 and 2020.

A much cited paper by Lotka (Lotka 1926) argued that mature scientific fields usually show a relationship between the number of authors contributing a single paper to a data set and the number of authors contributing to two or more outputs. This conjecture is usually referred to as Lotka's Law, expressed as

$$A = N/n^2$$

where N = the number of authors contributing to a single output;
n = the number of outputs an author contributes to;
and A = the expected number of authors contributing to this many papers.

Thus, if we have 439 authors who contribute to just one output, then we can expect to find 439/4 = 109 authors who contribute to two papers. The bottom line in Table 1 shows how many authors we would expect to find contributing to N papers. The table suggests that the L2 vocabulary research data for 2020 deviates quite markedly from Lotka's model. The field as a whole appears to be excessively reliant on authors who make a single contribution to the oeuvre. The proportion of authors contributing to a single output in this data is almost 90%—a considerably higher proportion than we noted in the earlier data. This feature is one that has been noted in my analyses of earlier data sets, but it is surprising to see it persisting in the 2020 data. In my earlier analyses, I suggested that the discrepancy might be a characteristic feature of a relatively new and relatively small research field, but the discrepancy reported here between Lotka's estimate and the actual publication data seems to be widening rather than decreasing as we might have expected. The reasons for this are unclear.

3. Counting the Citations

Obviously, the fact that an author publishes a lot of papers does not tell us whether this work is influential or not, but we can begin to assess the importance of authors' contributions to the field by examining the way their work is cited in the research literature. Most published work is cited only sporadically, but a few authors are widely cited by other researchers, and can be regarded as significant influences in the field. In this section, we move beyond the superficial analysis reported above and look at the citations found within a subset of the 2020 publications. This subset is made up of 238 journal articles and 65 book chapters. I extracted from these outputs a list of all the names who appeared as authors of papers cited in the data set. This process resulted in a list of 10,873 sources, an astonishingly large number. As usual, most of these sources are cited only once, but a small number of authors are cited many times in the data set. These data are summarised in Tables 2 and 3. Table 2 shows the number of sources who are cited N times in the 2020 data set. Table 3

lists the significant influences in the 2020 data set, along with the number of papers they are cited in.

Table 2. The number of cases who are cited N times in the 2020 data set.

No. Citations	1	2–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100	100+
No Cases	8142	2545	142	32	14	8	2	2	0	1	0	5

Table 3. The 20 most cited sources in the 2020 data set.

Source	Cited in N Papers	Source	Cited in N Papers
ISP Nation	206	R Waring	50
N Schmitt	173	A Coxhead	48
B Laufer	128	TS Paribakht	48
PM Meara	108	D Schmitt	48
S Webb	107	JL Milton	46
JH Hulstijn	84	A Pellicer-Sanchez	46
T Cobb	68	SD Krashen	45
J Read	65	M Horst	44
NC Ellis	59	W Grabe	40
D Beglar	53	MB Wesche	40

A number of points are worth making here.

Firstly, and most obviously, there is something of a mismatch between the list of Prolific Authors in the 2020 data, and the list of significant influences at this time. Only four authors figure in both lists (Coxhead, Laufer, Pellicer-Sanchez, and Webb). The best explanation for this is that the list of significant influences is biased towards authors who have published over a long period, and therefore have a lot of work that can be cited, while the list of Prolific Authors identifies active, younger authors whose work may not yet have had time to be cited. Laufer and Coxhead are the best counter-examples to this description: both of these authors have an extensive publication list and published three papers in 2020. Among the most-cited authors in 2020, several important figures have only two outputs in this year (notably Cobb, Nation, Schmitt, and Meara), while Read and Milton contributed to only one publication each. The remaining authors in the Most-Cited Sources list authored no papers in 2020, suggesting that their contribution to the field might be historical rather than on-going. The obvious case of this is SD Krashen, whose main work appeared in the 1990s, but is still strongly cited in the literature.

A second, but less obvious point is the extent to which "grade inflation" has become a significant factor in L2 vocabulary research. In the 1990 data (see Figure 1), only 118 relevant papers were identified, and it was relatively easy for an author to be identified as a significant influence. The most cited source at this time was Meara, who was cited in only 19 papers, 16% of the total output for that year. In contrast, 84 authors were cited at least 19 times in the 2020 data. Nation, the most cited author in the 2020 data, is cited in 206 papers—67% of the total output for that year—an astonishingly high figure, and more than 10 times the number of works that cited Meara in 1990. One hundred and twenty authors were cited in at least three papers in 1990, but most of them produced no L2 vocabulary-related outputs. In 2020, 462 sources were cited three or more times. This means that it is very much harder for an author to be identified as a significant influence in 2020 than it was in 1990.

A third point to make at this stage, one that we will take up again later, is that many of the significant influences in 1990 fail to appear in the Prolific Authors List and the Most-Cited Sources list in 2020.

4. The Co-Citation Analysis

So far, we have been discussing some relatively superficial characteristics of the 2020 data. In this section, we move onto a more detailed consideration of the patterns of co-citation within the data set. This kind of analysis can be traced back to a series of influential papers published by Small (e.g., Small 1973), suggesting that authors who are frequently cited alongside each other can help us to identify thematic trends in a body of research. In L2 vocabulary research, for example, one important strand of research concerns the structure of a bilingual speaker's mental lexicon, and recent research in this area will almost always cite pioneering work by de Groot, van Hell, Kroll, and Green. These four authors form the core of a cluster of researchers who tend to be co-cited in this context. The question we are interested in is what clusters can we identify in the 2020 research?

Author co-citation analysis, the main analysis tool used in this paper, traditionally makes a number of simplifying assumptions. The most straightforward assumption is that our analysis is based on two main types of research output: journal articles and chapters appearing in edited volumes. Monographs, theses, and non-standard reports are eliminated from the analysis on the grounds that the way references are used in work of this type tends to be very different from the way they are used in journal papers and similar reports. As we have seen in the previous section, the full 2020 data set contains seven books, three theses, and a computer program which are not included in the analysis that follows.

The second assumption traditionally made in author co-citation analyses is that the structure of the field is best captured by looking at a small number of sources who are most frequently cited in the data set. Conventionally, this number is set at about 100 sources—an arbitrary figure, largely driven by a need to keep the data manageable and easy to visualise.

The first step of the analysis involves identifying a set of authors who meet this inclusion criterion. With the 2020 data set, we can get close to this conventional figure by choosing to work with the 104 sources that are cited in at least 17 papers. (Note, though, that if we had adopted this criterion for the 1990 data set, only one source would have qualified—another example of the grade inflation that we noted in the previous section.)

The next step involves drawing up a large matrix which shows the number of times each of the 104 authors is co-cited with each of the other authors in the matrix. Once these data are in place, they can be submitted to a standard mapping program which converts the matrix to a two-dimensional mapping. In the analyses that follow, I used the *Gephi* program to perform this conversion (Bastian et al. 2009). Normally, this results in a mapping which looks like Figure 1, but, as we will see, the 2020 data set has some characteristics that make it difficult to process.

Figure 3 illustrates the main problem with this data set. This figure is a spanning tree which shows the strongest co-citation links between the most frequently cited 51 sources in the data set. Each of the 51 sources is cited at least 25 times in the data set. Each node is connected to the tree by its strongest co-citation link—the weakest link in the figure is the link between Nation and Kroll, which occurs only 14 times in the data set; all the other links are stronger than this. The ten strongest links in the spanning tree are listed in Table 4.

Table 4. The 10 strongest co-citation links in Figure 3. Note that all of these links involve Nation and one other source.

	Links	Weight		Links	Weight
1	Nation–Schmitt	156	6	Nation–Cobb	65
2	Nation–Laufer	122	7	Nation–Read	61
3	Nation–Webb	103	8	Nation–NC Ellis	50
4	Nation–Meara	101	9	Nation–Beglar	50
5	Nation–Hulstijn	79	10	Nation–Waring	48

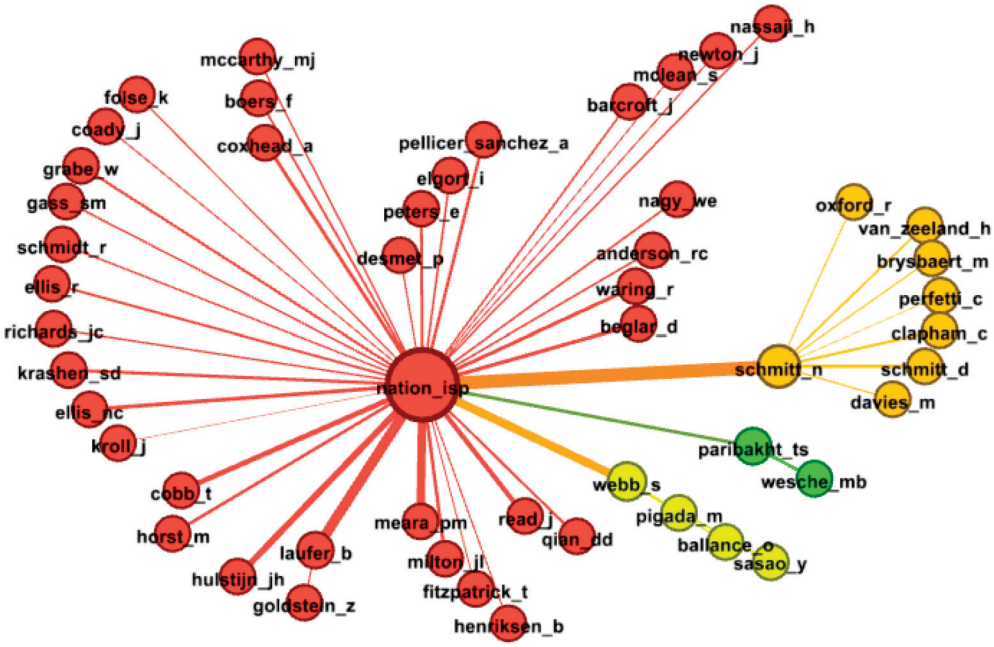

Figure 3. A spanning tree showing the 51 most-cited sources in the 2020 data set. The map shows the strongest connections between the nodes.

Gephi's analysis of these data suggests that we can identify four clusters in this data set—the clusters focussed on Nation, Schmitt, Webb, and Paribakht, respectively. However, this spanning tree mapping is very unusual, in that we have three very small clusters and one enormous cluster focussed on Nation. This large cluster includes 37 of the 51 most-cited sources in the 2020 data set—73% of all the sources that appear in the figure.[1] The most striking feature of this main cluster is that it has very little internal structure. Of the 37 members of this cluster, 36 members are maximally co-cited with Nation: the only source who does not fit this pattern is Goldstein (most often co-cited with Laufer). It is possible to identify some sub-clusters within the largest cluster, and I have arranged the nodes in Figure 3 to reflect a speculative interpretation of these sub-clusters. However, even where the focus of a sub-cluster is relatively clear—as with the sub-cluster whose main focus is on eye-tracking (Pellicer-Sanchez, Peters, de Smedt, and Elgort)—the sources are still more strongly co-cited with Nation than they are with each other. The smaller cluster focussed around Schmitt accounts for a further 16% of the sources, bringing the tally of sources counted in the two main clusters to 89% of the total.

Clearly, the 2020 map is structurally very different from the 1990 map shown in Figure 1, where the clusters are smaller. Here, Nation's influence is so dominant that it prevents the more detailed structure of the co-citations from becoming apparent. However, we can get around this problem by arguing that Nation's influence in the 2020 data is so all-pervasive that it does not actually make a distinctive contribution to the structure of the data set. This means that we might be able to ignore Nation, and explore the co-citation patterns in the rest of the data. Essentially, this involves building a "donut map" which ignores the massive impact of dominant sources like Nation, and allows weaker relationships to emerge from his shadow, as it were. Figure 4 shows the results of an analysis of this sort.

The donut map of the 2020 data set is fairly easy to interpret. The core of the map consists of six very significant influences. The predominant influence is Schmitt, followed closely by Laufer. Hulstijn, Webb, Cobb, and Meara make up the rest of this core, but the co-citations between all six of these influences are very strong. This core is surrounded by a group of lesser influences. Waring, Horst, Read, NC Ellis, Pellicer-Sanchez, Paribakht, Wesche, Beglar, Peters, Coxhead, and D Schmitt make up this group of less central authors. Their co-citation links are more focussed than those of the central core.

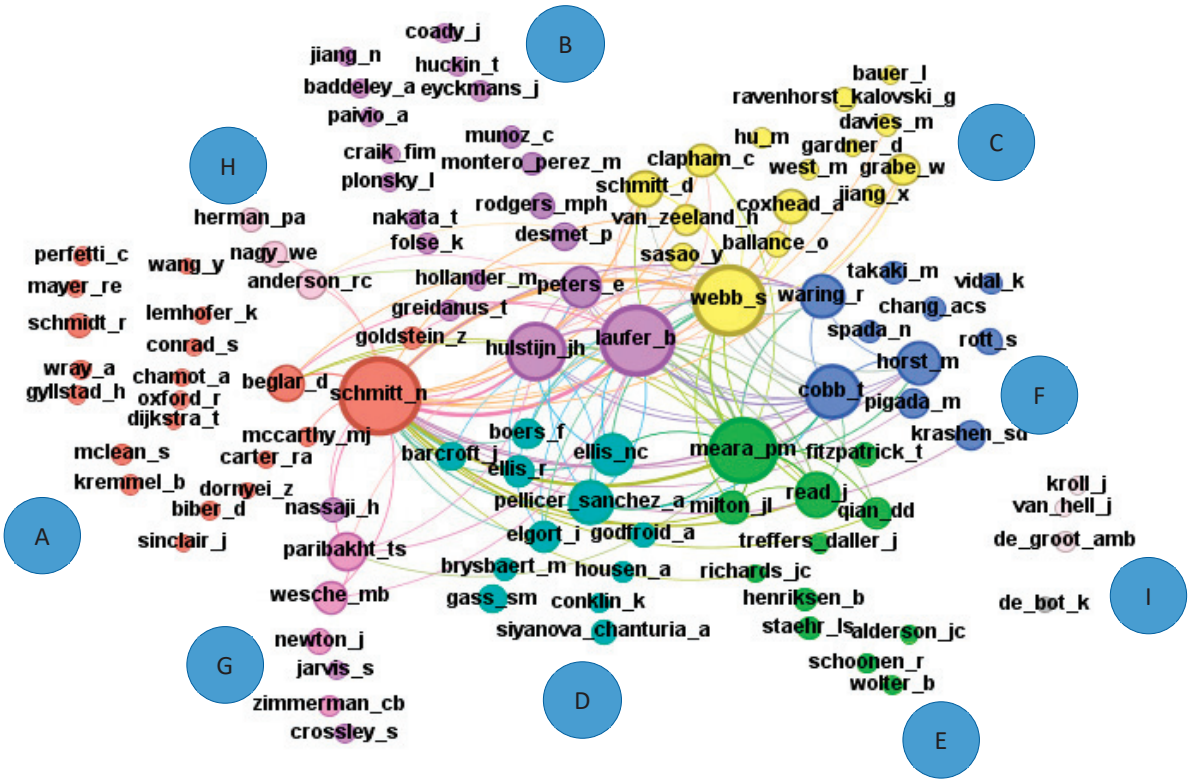

Figure 4. The 2020 data set without co-citations involving Nation. The figure contains 103 nodes, each co-cited at least 17 times on the data set. Edges with strength less than 25 have been removed in the interests of clarity. Nodes are sized according to the number of co-citation links they share in.

Gephi identified nine clusters in the 2020 data set, together with one isolated source (de Bot). These clusters are described below.

The largest cluster, **Cluster A**, is focussed on Schmitt. This cluster contains 20 members, many of whom are closely associated with Nottingham University. The cluster seems to include a number of sub-clusters. McCarthy, Carter, Biber, and Sinclair appear in my earlier analyses as a distinct cluster dealing with formal descriptions of lexis and corpus analysis. Chamot, Oxford, and Dornyei make up a sub-cluster that deals with motivation and strategies. Wray and Gyllstad represent a research strand that deals with collocations and multi-word lexical items. (Note, though, that the VARGA database does not systematically monitor research on multi-word items, so the analysis may be underestimating the importance of this research strand.)

Cluster B, with 19 members focussed on Laufer and Hulstijn, looks to be mainly concerned with implicit and explicit vocabulary acquisition in a variety of contexts.

Cluster C, with 14 members focussed on Webb, consists of a subgroup of sources working on word lists (Davies, Gardiner, Coxhead, and the historically important West), and the degree of coverage provided by these word lists. It also includes Diane Schmitt and Clapham, whose work with Norbert Schmitt provides a standard assessment tool for measuring vocabulary mastery that is based on frequency counts.

Cluster D, with 12 members focussed on Pellicer-Sánchez, seems to be a methodologically motivated cluster, whose members use eye-tracking to study L2 reading.

Cluster E, also with 12 members, is also a characterised by a distinctive methodology. This cluster seems to be mainly concerned with lexical inferencing, but it also contains a sub-cluster that uses word-association methods in an attempt to assess vocabulary depth. This cluster is particularly associated with the research group at Swansea University.

Cluster F, with 10 members focussed on Cobb, is also concerned with vocabulary uptake from reading, though Cobb's web-site is an important methodological resource in a number of disparate areas.

Cluster G, a smaller cluster with only seven members, is another methodological cluster, this one concerned with ways of assessing change in vocabulary knowledge. Wesche and Paribakht's Vocabulary Knowledge Scale is the standard tool here (cf. Wesche and Paribakht 1996).

Cluster H, with only three members, is a group of L1 reading researchers whose work has seriously influenced the thinking of L2 vocabulary research.

Cluster I, again with only three members, is the remains of a psycholinguistics research tradition dealing with the structure of bilingual lexicons. This tradition played a very significant role in earlier periods of L2 vocabulary research, but by 2020 this role seems to be diminishing in this map, and the members of this cluster have become detached from the main L2 vocabulary research clusters. There are some weaker links with other nodes, but they are not strong enough to appear in Figure 3.

De Bot appears in Figure 3 as a detached singleton with no connections to the rest of the network. This, again, is the result of the filters applied to the data: de Bot is substantially cited in the 2020 data set (17 times), but he is most frequently co-cited with authors who do not appear in the list of most-cited sources.

A number of points emerge from this analysis. Firstly, only nine significant influences survived from 1990 through to 2020. All nine sources were still co-cited in the 2020 data set, and some were very strongly co-cited (See Figure 5). The core of this map is the Laufer–Meara–Nation triangle. Anderson and Nagy are part of an L1 reading theme, while Krashen is mainly cited for his claims about incidental L2 vocabulary acquisition from reading. McCarthy is the main survivor from the corpus linguistics theme that dominated L2 vocabulary research in the early 1990s. Richards is mostly cited for his classic work on what it means to know a word (Richards 1976). Gass is mostly cited here for her recent work on subtitled input and eye-tracking.

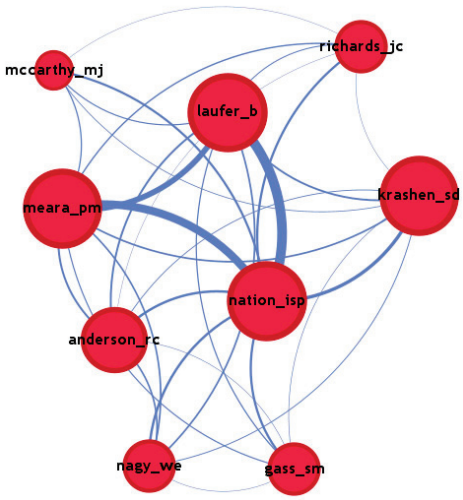

Figure 5. The nine significant influences who appear in both the 1990 data set and the 2020 data set.

Clearly, the majority of the significant influences who appear in the 2020 map are relatively recent new-comers (See Table 5). I do not yet have enough data to determine exactly when these new sources rose to prominence, but it is probably safe to say that Schmitt, Hulstijn, Read, and Cobb first become important sources in the 1990s, while most of the other new sources have a much shorter bibliometric history.

Table 5. The most significant influences in the 1990 data set and the 2020 data set.

1990 only	Aitchison, JR Anderson, R Brown, Carter, Channell, AD Cohen, Corder, Elley, Faerch, Feldman, Francis, Gairns, Haastrup, Jain, Kasper, Kellerman, M King, Kirsner, Levenston, Lockhart, Ostyn, Palmberg, Putseys, Redman, Rudzka Schouten_van Parreren, MC Smith, Swain, Wallace, West
Both years	RC Anderson, Gass, Krashen, Laufer, McCarthy, Meara, Nagy, Nation, Richards;
2020 only	Barcroft, Beglar, Boers, Brysbaert, Clapham, Cobb, Coxhead, Desmet, Elgort. NC Ellis. R Ellis, Grabe, Horst, Hulstijn, Milton, Nagy, Nassaji, Oxford, Paribakht, Pellicer-Sanchez, Peters, DD Qian, J Read, Sasao, Schmidt, N Schmitt, van Zeeland, Waring, Webb, Wesche

The second point to emerge from the analysis is that a number of significant influences who appeared in the 1990 data set can no longer lay claim to this role in the 2020 data. Some of these authors continue to be cited in the 2020 data, but their influence is a pale shadow of what it was. Corder, Schouten-van Parreren, Palmberg, and Levenston—all very significant influences in the 1990 data set—are cited only once in 2020, for example. Kellerman and Stevick are cited only twice. An interesting case is Aitchison. She was cited five times in the 1990 data set. In the 2020 data set she is cited 14 times: far more frequently than she was cited in 1990 but slightly fewer times than the stricter threshold I have used for the 2020 map (another example of the grade inflation trend that I mentioned earlier.)

This leads us to ask what areas of interest present in the 1990 map no longer have a presence in the 2020 map. I think we can identify six main areas that were important in 1990, but seem to be less important in 2020. The most obvious loss is the disappearance of all the members of Cluster III, Cluster IV, and Cluster VI in the 1990 map (see Figure 1). These clusters contain a number of sources whose main interest was the practical applications of vocabulary theory to teaching—Gairns and Redman, and Rudzka, Ostyn, Channell, and Putseys are particularly important in this context. Also important is the disappearance of the Scandinavian research group—Palmberg, Phillipson, Haastrup, Faerch, and Kasper—that strongly influenced the 1990s research on lexical inferencing. Transfer and lexical errors (Kellerman, Ringbom, and Corder) no longer seem to play a role in the 2020 research. L1 vocabulary development (HH Clark and R Brown) was an important influence in 1990 but not, it seems, in 2020. None of the members of the psycholinguistics cluster (Cluster II in Figure 1) met the inclusion threshold for the 2020 map, although this strand of research is represented by three new sources (de Groot, van Hell, and Kroll) in 2020.

What replaces these historical concerns? Figure 6 shows a spanning tree analysis of the new sources in the 2020 map, and we can begin to answer this question by identifying the clusters which emerge in this map.

This map, again, is relatively easy to interpret. The spine of the map runs from left to right, from Hulstijn through Schmitt to Webb and Cobb. Gephi found 10 clusters in this data set, but once again we have the problem that a single node dominates the map and prevents the finer structure from emerging. Schmitt and Webb seem to play the same role here as Nation did in Figure 3. Gephi did manage to find some interpretable clusters in this tree, but these clusters are mostly formed of sources who co-author a small number of frequently cited papers (Wesche and Paribakht 1996; Hulstijn et al. 1996; and Cobb and Horst 2019, for example). It is possible to disaggregate the single massive cluster dominated by Schmitt and Webb, and I have reflected this in the way I have drawn the map. However, the dominant position of Schmitt suggests that what we have here is another case where a donut map makes more sense than the straightforward spanning tree. Figure 7 shows a donut map of the newcomers with Schmitt's co-citations and Webb's co-citations excluded from the analysis.

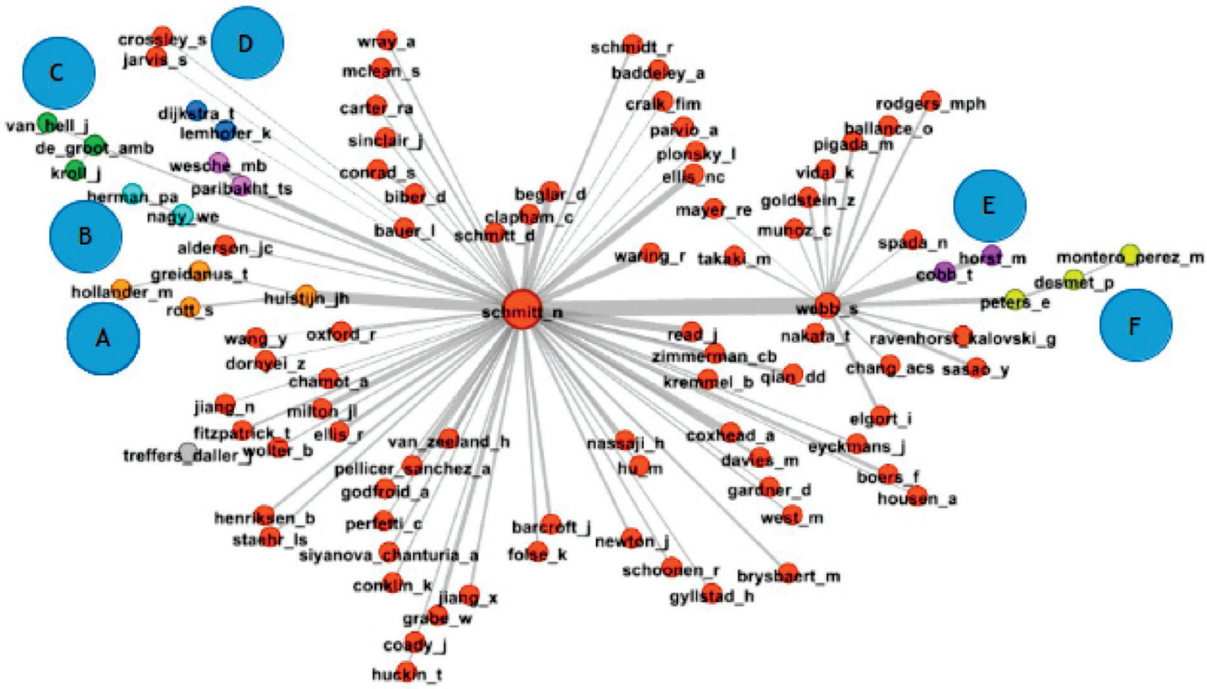

Figure 6. A spanning tree analysis of the new sources who appear in the 2020 map. Each source is connected to the source it is most frequently co-cited with.

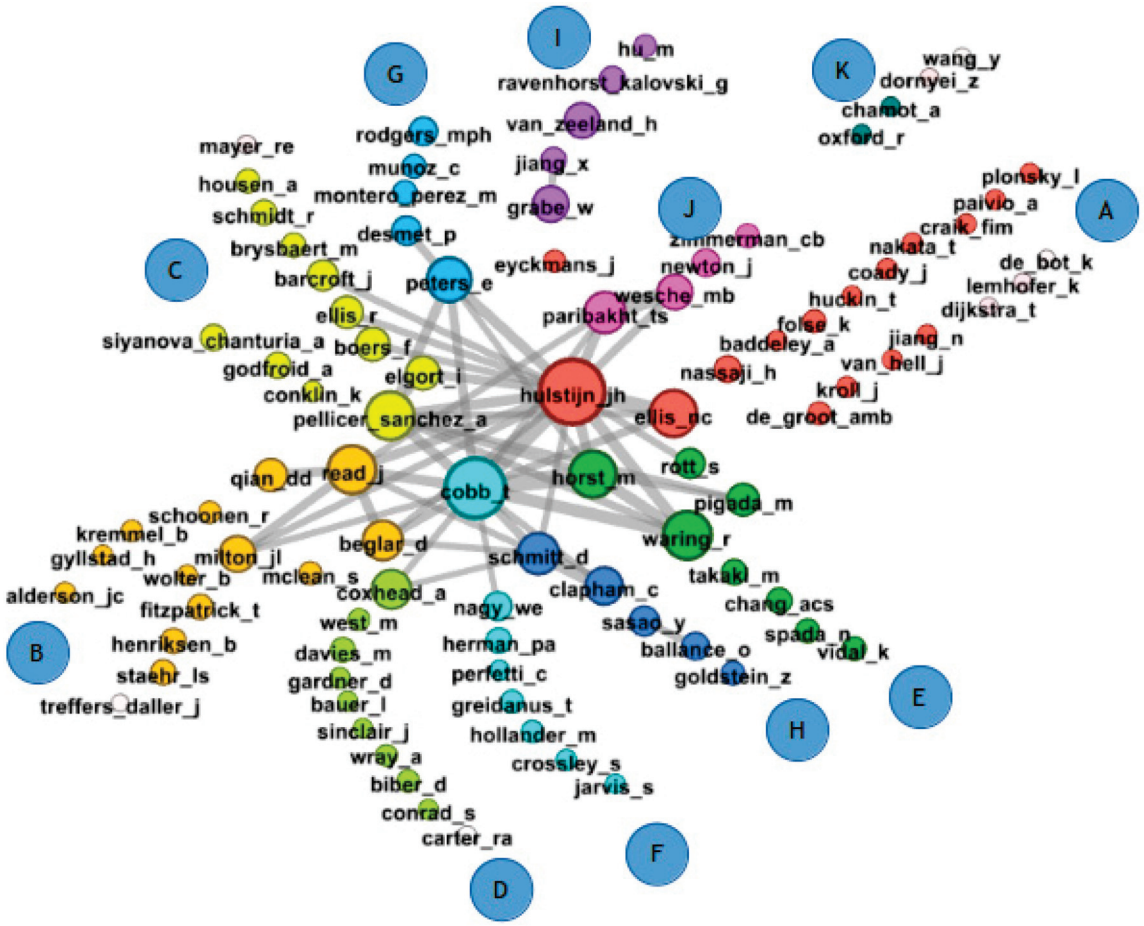

Figure 7. The new entries in the 2020 data set: donut mapping with Schmitt and Webb excluded. Threshold for inclusion: 25 co-citations. Only edges with weight greater than 20 are shown.

This mapping is extremely easy to interpret. The core of the map is a set of highly interconnected hubs. The most significant influence in the map is Hulstijn, followed closely by NC Ellis, Cobb, Horst, Read, Pellicer-Sanchez, and Peters. The strongest co-citation links between these central hubs are listed in Table 6.

Table 6. The 15 strongest co-citation links in Figure 7.

	Links	Weight		Links	Weight		Links	Weight
1	Cobb–Horst	40	**6**	Cobb–Waring	31	**11**	Cobb–Read	27
2	Paribakht–Wesche	39	**7**	Hulstijn–Waring	29	**12**	Cobb–Pellicer-S.	27
3	Cobb–Hulstijn	38	**8**	Horst–Waring	29	**13**	Beglar–Read	27
4	NC_Ellis–Hulstijn	37	**9**	Hulstijn–Peters	28	**14**	Hulstijn–Paribakht	27
5	Clapham–D_Schmitt	35	**10**	Horst–Hulstijn	27	**15**	Peters–Pellicer-S.	27

Gephi's analysis suggests that there are 11 clusters in this data set, and they can be straightforwardly identified as the main research trends in the 2020 data set.

Cluster A, the largest cluster focussed on Hulstijn and NC Ellis, seems to consist of two smaller clusters, one comprised mainly of psychologists whose work has strongly influenced research on L2 word recognition, the other dealing more specifically with word recognition by bilinguals.

Cluster B, focussed on John Read, looks like a group of sources interested in depth of vocabulary knowledge. This cluster also contains a sub-cluster that focusses on word-associations.

Cluster C, focussed on Pellicer-Sanchez, contains a number of sources who use eye-tracking as their main methodology.

Cluster D, focussed on Coxhead, is largely concerned with word lists and analyses of corpora for L2 vocabulary teaching.

Cluster E, focussed on Horst and Waring, is mainly concerned with vocabulary acquisition from reading.

Cluster F is strongly influenced by the coh-metrix approach to learner output.

Cluster G is concerned with modified input for L2 learning.

Cluster H is mainly focussed on measures of vocabulary size.

Cluster I is an L2 reading cluster.

Cluster J is a set of sources that uses Wesche and Paribakht's Vocabulary Knowledge Scale as a way of assessing the partial learning of words.

Finally, **Cluster K**, focussed on Oxford, is a set of sources whose main interest is learning strategies.

Most readers who are familiar with the current L2 vocabulary research would probably agree that this map does capture the main trends in the research quite well. However, we need to bear in mind that a lot of data were eliminated in order for us to reach this plausible conclusion. See Section 5 below.

An alternative, more nuanced conclusion is that the current research can be seen to be made up of four main components: we have a core of historically important sources closely associated with Nation; a more recent core of very significant sources (Schmitt and Webb) that is very widely co-cited in the more recent research but not in the earlier research; a number of new sources that act as hubs for a cluster; a number of other, but less influential, sources. This last group can be split into two smaller groups: a group of older sources that were significant in earlier maps, whose influence appears to be waning (Corder and Krashen are good examples of this), and a small group of genuinely new sources whose importance is on the rise.

What the maps cannot tell us is what kind of temporal trajectory is being followed by these sources. Clearly the maps do not tell us anything in detail about what is going on between the two snapshots taken in 1990 and 2020. Some of the research trends identified in the maps are recent and genuinely new (eye-tracking, for example). Others are long-standing interests, but the sources that allow them to be identified have changed (de Groot,

van Hell, and Kroll, for example, replacing Kirsner and colleagues investigating the way bilingual lexicons are organised). The thirty-year gap between 1990 and 2020 is very large in research terms, covering several generations of research students. Clearly, more work is needed to clarify how the field is developing between these two snapshots (Meara 2020).

5. The Other Research

So far, I have discussed the broad patterns of author co-citations in the 2020 data set, with a particular focus on the role of Nation in shaping the overall structure of this research. I have also noted that this discussion was based on an analysis of the most frequently cited 104 authors in the 2020 data. It will be obvious, however, that restricting the analysis in this way means that we are not taking any account of the many other sources that are not cited to the same extent. Figure 7, for example, takes no account of the sources in the 2020 data set that are cited fewer than 17 times. The data set actually contains 2601 sources that are cited more than once, so the mapping in Figure 7 is based on only 4% of the data available for analysis. In contrast, the 1990 map shown in Figure 1 captured nearly 23% of all the co-citations that appeared twice or more in the data set, and this makes it much easier to justify the inclusion threshold—only the very weakest of co-citation links were ignored in the 1990 map. Clearly, this problem becomes more serious as the data sets get larger, and the inclusion threshold that gives us about 100 sources excludes an ever-increasing proportion of the data set.

This difficulty seems to arise because of the grade inflation factor. A particular factor is that the average number of research outputs cited by a typical paper has exploded since 1990. In 1990, the average number of sources cited in a paper was 25—though many papers cited fewer than 10 sources. By 2020, this figure had increased to 74, with 17 papers citing more than 150 authors, and one paper (Sulpizio et al. 2020) citing 673 authors! Partly, of course this increase comes about simply because there is more research available to be cited, a significant increase in the number of papers with multiple authors, and the increasing appearance of meta-analyses, but this is not the whole story. As the number of research outputs increases, the co-citation statistics are dominated by derivative research, papers that largely duplicate things we know already, do not use innovative methodologies, and do not push the boundaries of theory. This type of research tends to cite "the usual suspects", sources that lend a certain amount of legitimacy to a paper, but are not necessarily engaged with. The result is that the co-citation maps become overwhelmed by a canonical list of sources that are widely cited, but are not truly reflective of the current research fronts. These features do not entirely invalidate the analysis that I have presented in this paper, but they do suggest that research fronts in L2 vocabulary research might be more elusive and harder to recognise than we had expected.

This raises the question whether the conventional practice of working with only the most frequently cited sources in a data set, producing mappings which take into account only a hundred of these sources, needs to be rethought. Perhaps we should be looking for research clusters that meet common objective criteria, rather than the arbitrary thresholds that underpin maps like Figure 7?

I have not yet managed to find a methodology that could automatically identify and extract less frequently cited but still interesting clusters in very large data sets. However, anyone familiar with the entire 2020 oeuvre will realise that there are a number of themes running through this work which look as though they might be embryonic themes—themes that are not cited often enough for them to appear in the main mappings, yet nonetheless capture some important growing topics of interest in L2 vocabulary research. I have listed some of these themes in Table 7, along with some papers published in 2020 which exemplify them. Several of these themes represent a substantial amount of research output: the gaming theme, for instance, appears in at least 10 papers in 2020. Gaming does not appear explicitly in Figure 7. Had these papers appeared in 1990, however, they would easily have met the threshold for inclusion in the 1990 map, and gaming would have appeared in Figure 1 as a very significant research cluster.

Table 7. A set of topics which frequently occur in the 2020 data set, but fail to meet the threshold for inclusion in the main analysis.

Topic	Examples in the 2020 Outputs
Technology and apps	(Enayati and Gilakjani 2020); (Tai et al. 2020); (Obermeier 2020)
Data-driven approaches	(Talib 2020); (Lee et al. 2020); (Dang 2020)
Listening vocabulary	(Masrai 2020); (Zhang and Graham 2020)
Gestures	(Andrä et al. 2020)
Vocabulary Networks	(Aotani and Takahashi 2020); (Spätgens and Schoonen 2020)
CLIL	(Baten et al. 2020); (Benatti 2020)
Neurolinguistic Approaches	(Zhang et al. 2020); (Zeller 2020); (Liu and Hell 2020)
Gaming approaches	(Bahari 2020); (Thompson and Gillern 2020)
Trilingual speakers	(Chen and Liu 2020); (Efeoglu et al. 2020); (Gudmundson 2020)
Young learners	(Alexiou and Milton 2020); (Nordlund and Norberg 2020)
Availability	(Akbarian et al. 2020)
Learner Autonomy	(Denison and Custance 2020); (Kambakis-Vougiouklis and Katsarou 2020)
Spaced learning	(Sararoodi and Farvardin 2020); (Lafleur 2020)
Good vocabulary learners	(Sasao 2020); (Nahavandi et al. 2020)
Teaching strategies	(Herrera Caldas et al. 2020)

Space does not allow me to follow up this idea here, and we cannot investigate the characteristics of these sub-threshold clusters in any detail. However, I will be reporting an analysis that tries to establish objective criteria for identifying clusters of this sort in a follow-up paper. In the meantime, there is one set of research papers which can be objectively identified and analysed using the same approach that we used in the previous section. This is a subset of the 2020 data set that does not cite Nation among its sources. This work is not marginal, but its relationship to the work that does cite Nation means that it tends to get overlooked. It is instructive to examine how this work differs from the work we have discussed so far.

The 2020 data set contains a subset of 74 papers which unusually do not cite Nation among their references. These 74 papers do cite other sources, however—5214 in total, of which 4468 are cited only once. The distribution of these papers is summarised in Table 8. The most frequently cited sources in this subset of the data are Schmitt (19 citations); Kroll (12 citations); van Hell (11 citations); and Bates, Brysbaert, Laufer, and Perfetti (10 citations each). The overall distribution of these citations is shown in Table 9. What is striking here is the fact that only two of the very significant influences that we identified in the larger 2020 data set appear in this smaller set (Schmitt and Laufer).

Figure 8 shows this smaller data set as a spanning tree mapping. The map is made up of 61 sources that are co-cited at least five times in the reduced data set; each node is connected by an edge to the node it is most frequently co-cited with. The spine of this mapping is the set of co-citations that link van Hell, Kroll, Bates, Brysbaert, and Schmitt. The strongest links in the mapping are reported in Table 9.

Gephi's analysis found eight clusters in this data set.

Cluster A, dominated by van Hell, is made up of sources that are interested in bilingual lexicons, and use word recognition as their main methodological tool. **Cluster B**, dominated by Kroll, shares many of the same concerns as Cluster A, but seems to be more concerned with productive vocabulary, whereas Cluster A focusses on receptive vocabulary. Cluster B also makes use of more formal models, notably the Bilingual Interactive Activation model (Dijkstra et al. 1998). **Cluster C** is a set of statistical sources that identify methodological features in Cluster A and Cluster B. Cluster D, dominated by Brysbaert, also seems to be a methodological cluster. Brysbaert has published a large set of vocabulary size tests for several different languages, and a number of norms lists which are routinely used by the sources in Cluster A and Cluster B. The other members of this cluster seem to be interested mainly in reading behaviour. This theme is also picked up in **Cluster F**, a group of sources whose main interest is reading behaviour in young bilingual speakers. **Cluster E** is a much reduced set of the influences who appear in Figure 6. The striking feature here is how small

this cluster is, and how few co-citation links there are between this cluster and the rest of the map. **Cluster F** is a small, detached cluster of Canadian researchers who mainly work on English/French bilinguals. The detached status of this cluster starkly emphasises the dominance of English language research in all the maps we have discussed so far in this report.

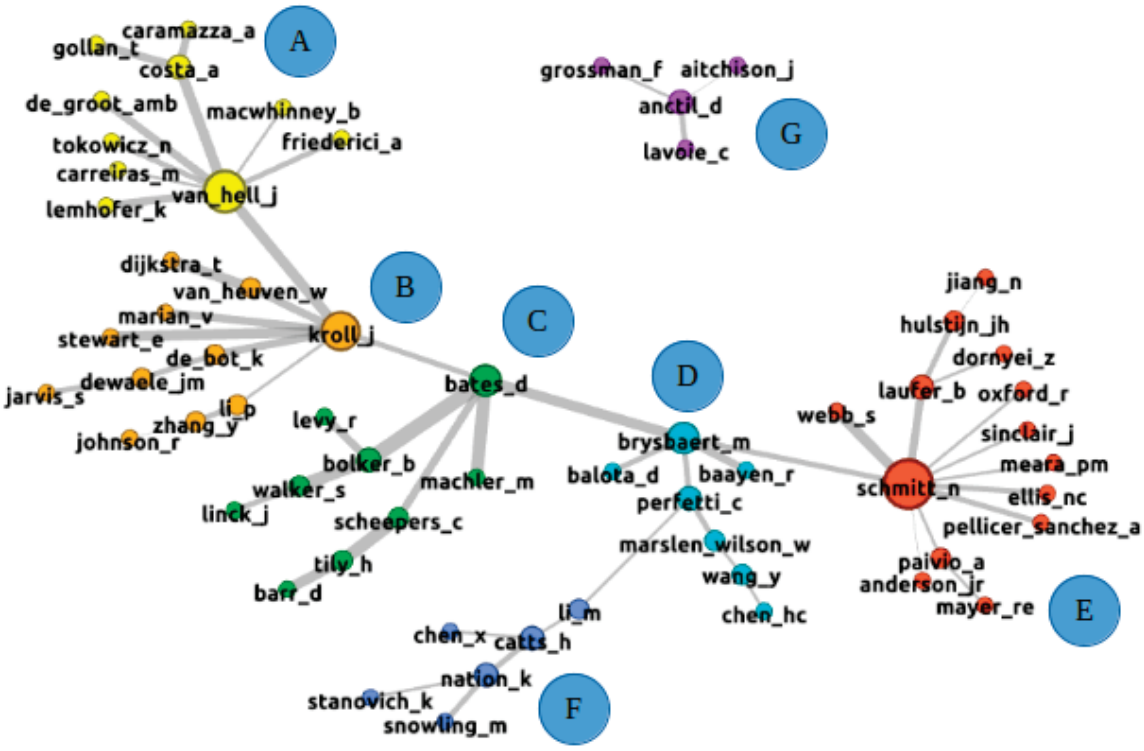

Figure 8. A co-citation map based on the 61 sources in the 2020 data set that do not cite Nation. Each source is cited at least five times in the data set.

Table 8. The number of cases who are cited N times in the not-Nation 2020 data set.

No. Citations	**20**	**19**	**18**	**17**	**16**	**15**	**14**	**13**	**12**	**11**
No. Cases		1							1	1
No. Citations	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
No. Cases	4	3	4	9	6	32	55	152	478	4468

Table 9. The strongest co-citation links in Figure 8.

	Links	**Weight**		**Links**	**Weight**		**Links**	**Weight**
1	Bates–Bolkert	8	**6**	Kroll–Stewart	6	**11**	Schmitt–Webb	6
2	Bates–Machler	7	**7**	Costa–van Hell	6			
3	Scheepers–Tily	7	**8**	Kroll–van Hell	6			
4	Bolkert–Walker	6	**9**	Bates–Brysbaert	6			
5	Barr–Tily	6	**10**	Dijkstra–v Heuven	6			

Almost all of the sources identified in Figure 8 will be unfamiliar to researchers working in the more mainstream L2 vocabulary tradition. Figure 8 strongly suggests that the research clusters identified in Figure 7 are only part of a much more complicated research endeavour. Yes, the main bulk of the 2020 research clusters around a small number of very significant sources, that most researchers in the Applied Linguistics tradition would immediately recognise, but there exists alongside these sources at least one other "invisible college", a set of clusters which is numerically large, intellectually coherent, but to a large

extent independent of the main research trends identified in Figure 7. The data suggest that we have two quite separate research traditions here, one where researchers see their roots in Applied Linguistics research, the other looking more towards psychology. There is very little direct contact between these two traditions, with almost no overlap in the sources that they cite, despite a shared set of research questions, particularly an interest in L2 reading behaviour. There is clearly a case for encouraging closer collaboration between the two traditions.

6. Conclusions

This paper has presented an interim analysis of the L2 vocabulary research published in 2020. The introduction to this paper raised a number of questions about how L2 vocabulary research developed between 1990 and 2020. Not all these questions were answered by the data that I have reported so far, however. The two snapshots discussed suggest that there have been some major changes between 1990 and 2020—particularly the enormous growth in research outputs, the huge increase in the number of researchers active in the field, and the appearance of several new research themes. With only two data points, it is not possible to present the kind of fine-grained analysis that we would really like. Work of this type requires much more data than are currently available. In broad terms, however, we have noted some major changes in the field. Almost all of the authors who appear in the 1990 map have been replaced by a new generation of researchers, and some well-established figures no longer appear to be influential in 2020. The field as a whole is still driven by practical concerns (making L2 vocabulary learning effective, and research into teaching methods), but much of this work is not cited often enough for it to appear in the bibliometric maps. Most of the clusters identified in the 2020 map were also present in the 1990 map, but some have become less central in the intervening years: the distinctive Scandinavian approach to vocabulary acquisition that we identified in the 1990 map seems to have disappeared, and the concern with word frequency identified by cluster IX in Figure 1 no longer seems to be important in the more recent maps. The Carter/McCarthy corpus sub-cluster in Figure 1 seems to have morphed into to a more general concern with phraseology (Cluster D in Figure 7). The more recent maps show a heightened use of standardised research tools (notably Wesche and Paribakht's Vocabulary Knowledge Scale and the several re-workings of Nation's Levels Tests). Some methodological innovations are also apparent—notably, the appearance of eye-tracking research. An important recent development is the appearance of a lot of neuro-linguistic research that uses clinical methods such as fMRI scanning and event-related potential (ERP) studies. This work is not yet cited often enough for it to appear as a separate cluster in the significant influences map, but it is clearly a trend that will play a large role in future maps.

The main characteristic of the 2020 data set is the dominant position of Paul Nation: Nation is particularly strongly cited by authors whose main interest lies in vocabulary teaching, but not exclusively so. Among the top 104 most frequently cited authors, nearly all are co-cited with Nation. On the other hand, we also have a small number of authors who do not cite Nation in their own papers, notably sources working on L2 word recognition and some aspects of L2 reading. These sources have all the characteristics of an "invisible college"—they make up a rich, strongly structured set of influences with a distinctive approach to L2 vocabulary acquisition. However, they do not have much of a presence in our main analysis of the 2020 data set, and they are not easily identifiable as a thematic cluster in Figure 6.

This is not a new phenomenon. Most of my earlier maps—covering research before 1990—included a cluster of sources that would probably identify themselves as psychologists rather than linguists, and there has always been a tendency for these sources not to be co-cited with mainstream L2 research. Sadly, there have been few attempts to bridge the growing gap between these two traditions. (One early, but largely unsuccessful attempt at bridge building can be found in Schreuder and Weltens 1993). In the 2020 data set, it looks as though the psycholinguistic theme in L2 vocabulary research is continuing to

develop as an independent research tradition which has only tenuous connections with the mainstream Applied Linguistics approach to L2 vocabulary research. Given that both traditions are concerned with a set of common interests, this cannot be a good thing, and there must be a strong case for a closer collaboration in future.

One particularly important characteristic of the psycholinguistic research is the way that it uses formal models of L2 lexical processing. Mainstream Applied Linguistics L2 vocabulary research relies much more heavily on informal, metaphorical models, such as Aitchison's "gigantic multi-dimensional cobweb" metaphor (Aitchison 1987, p. 72). Serious modelling is almost completely absent from the mainstream Applied Linguistics research, but the maps I have presented here suggest that modelling approaches might provide a way to inject some important new ideas into the field by forging new co-citation links between otherwise disparate research clusters.

Finally, this paper has pointed out some problems that appear when we try to conduct historical research using co-citation methods. "Grade inflation" over a 30-year period is a serious issue which makes it difficult to draw direct comparisons between work produced in the 1990s and more recent research. It is also unclear whether some of the operating assumptions developed by earlier researchers can be easily applied to the much larger data sets that are typical in more recent research. In particular, we noted that the practice of drawing maps based on the most-cited 100 sources becomes increasingly contentious when more and more sources are ignored because an arbitrary inclusion threshold has been adopted. I think this problem might be resolved by developing a set of objective criteria for identifying clusters in the citation data sets, but work of this sort lies beyond the scope of this paper. Much the same might be said for the spanning tree maps that I have used in this report. These trees are easy to read and interpret, but this simplicity is achieved at a cost: the spanning trees display only the very strongest links in the data set, and they do not display the weaker links that exist between sources. This means that they tend to prioritise the status quo rather than the cutting edge of research. Spanning trees may not be a good solution to the problem of how to map the complexities of a rapidly developing field rather than a static one. Again, a solution to this problem lies outside the scope of this paper. I would be happy to share my extensive data sets with any colleagues who can think of a better solution.

Funding: This research received no external funding.

Institutional Review Board Statement: This research did not involve human beings or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no relevant conflicts of interest.

Note

1 Sharp-eyed readers will notice that I have included only 51 sources in this mapping, rather than the conventional 100. I did try to draw convincing maps using 104 nodes, but almost all the additional nodes were most strongly co-cited with Nation, and the resulting tree was almost impossible to read.

References

Aitchison, Jean. 1987. *Words in the Mind*. Oxford: Blackwell.

Akbarian, Is'haaq, Fatemeh Farajollahi, and Rosa Maria Jiménez Catalán. 2020. EFL learners' lexical availability: Exploring frequency, exposure, and vocabulary level. *System* 91: 1–12. [CrossRef]

Alexiou, Thomai, and James Milton. 2020. Pic-Lex: A new tool of measuring receptive vocabulary for very young learners. In *Advancing English Language Education*. Edited by Wafa Zoghbor and Thomaï Alexiou. Dubai: Zayed University Press, pp. 103–13.

Andrä, Christian, Brian Mathias, Anika Schwager, Manuela Macedonia, and Katharina von Kriegstein. 2020. Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. *Educational Psychology Review* 32: 815–50. [CrossRef]

Aotani, Noriko, and Shin'ya Takahashi. 2020. An analysis of Japanese EFL learners' lexical network changes. *Frontiers in Education* 5. [CrossRef]

Bahari, Akbar. 2020. Game-based collaborative vocabulary learning in blended and distance L2 learning. *Journal of Open, Distance and e-Learning* 35: 1–22. [CrossRef]

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media* 3: 361–62. [CrossRef]

Baten, Kristof, Silke van Hiel, and Ludovic de Cuypere. 2020. Vocabulary development in a CLIL context: A comparison between French and English L2. *Studies in Second Language Learning and Teaching* 10: 307–36. [CrossRef]

Benatti, Ruben. 2020. CLIL courses: Teaching Italian language and culture in Turkmenistan. In *Current Perspectives on Vocabulary Learning and Teaching*. Edited by Nuray Alagozlu and Vedat Kiymazarslan. Newcastle: Cambridge Scholars Publishing, pp. 161–79.

Chen, Jianlin, and Hong Liu. 2020. The effect of the non-task language when trilingual people use two languages in a language switching experiment. *Frontiers in Psychology* 11. [CrossRef] [PubMed]

Cobb, Tom, and Marlise Horst. 2019. Bringing home the Word. *Canadian Modern Language Review* 75: 285–98. [CrossRef]

Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.

Dang, Thi Ngoc Yen. 2020. Corpus-based word lists in second language vocabulary research, learning and teaching. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. Abingdon: Routledge, pp. 288–303.

Denison, Clint, and Imogen Custance. 2020. Vocabulary learning using student-created class vocabulary lists. *Vocabulary Learning and Instruction* 9: 1–8. [CrossRef]

Dijkstra, Ton, Walter van Heuven, and Jonathan Grainger. 1998. Simulating cross-language competition with the Bilingual Interactive Activation Model. *Psychologica Belgica* 38: 177–96. [CrossRef]

Efeoglu, Gulumser, Gulru Yukesel, and Suat Baran. 2020. Lexical cross-linguistic influence: A study of three multilingual learners of L3 English. *International Journal of Multilingualism* 17: 535–51. [CrossRef]

Enayati, Fatemeh, and Abbas Pourhosein Gilakjani. 2020. The impact of computer assisted language learning (CALL) on improving intermediate EFL Learners vocabulary learning. *International Journal of Language Education* 4: 96–112. [CrossRef]

Gudmundson, Anna. 2020. The mental lexicon of multilingual adult learners of Italian L3: A study of word association behavior and cross-lingual semantic priming. In *Third Language Acquisition*. EuroSLA Studies 3. Edited by Camilla Bardel and Laura Sanchez. Berlin: Language Sciences Press, pp. 67–110. [CrossRef]

Herrera Caldas, Veronica, Erzsebet Bekes, and Carmen Cajamarca. 2020. Effective vocabulary acquisition strategies employed by Ecuadorian teachers. *Asian Journal of English Language Studies* 8: 159–88.

Hulstijn, Jan, Merel Hollander, and Tine Greidanus. 1996. Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use and the reoccurrence of unknown words. *Modern Language Journal* 80: 327–39. [CrossRef]

Kambakis-Vougiouklis, Penelope, and Eirene Katsarou. 2020. Exploring the role of self-regulation capacity and self-esteem on vocabulary learning strategy use by Greek university learners. *Asian EFL Journal* 27: 123–46.

Lafleur, Louis. 2020. The indirect spaced repetition concept. *Vocabulary Learning and Instruction* 9: 9–16. [CrossRef]

Lee, Hansol, Mark Warschauer, and Jang Ho Lee. 2020. Toward the establishment of a data-driven learning model: Role of learner factors in corpus-based second language vocabulary learning. *The Modern Language Journal* 104: 345–62. [CrossRef]

Liu, Yushuan, and Janet van Hell. 2020. Learning novel word meanings: An ERP study on lexical consolidation in monolingual, inexperienced foreign language learners. *Language Learning* 70: 45–74. [CrossRef]

Lotka, Alfred J. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16: 317–24.

Masrai, Ahmed. 2020. Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review* 11: 423–47. [CrossRef]

Meara, Paul. 2020. The emergence of a First Paradigm in vocabulary research: The bibliometrics of System. *Vocabulary Learning and Instruction* 9: 1–32. [CrossRef]

Meara, Paul. 2022. Ground Zero: A bibliometric analysis of L2 vocabulary research 1986–1990. *Linguistics Beyond and Within* 8: 133–51. [CrossRef]

Meara, Paul. 2023. Shifting sands: A bibliometric analysis of L2 vocabulary research in 1991. *Linguistics Beyond and Within* 9: 112–32. [CrossRef]

Meara, Paul. n.d. VARGA: The Vocabulary Acquisition Research Group Archive. Available online: https://www.lognostics.co.uk/varga/ (accessed on 20 December 2023).

Nahavandi, Zohreh, Mona Tabatabaee-Yazdi, and Aynaz Samir. 2020. Verbal and non-verbal fluid intelligence as predictors of vocabulary knowledge. *Journal of English Language Research* 1: 12–21.

Nordlund, Marie, and Cathrine Norberg. 2020. Vocabulary in EFL teaching materials for young learners. *International Journal of Language Studies* 14: 89–116.

Obermeier, Andrew. 2020. Exploring the effectiveness of deliberate computer-assisted language learning. *Vocabulary Learning and Instruction* 9: 24–38. [CrossRef]

Price, Derek J. de Solla. 1965. Networks of Scientific Papers. *Science* 149: 510–15. [CrossRef] [PubMed]

Richards, Jack C. 1976. The role of vocabulary teaching. *TESOL Quarterly* 10: 77–89. [CrossRef]
Sararoodi, Ashraf, and Mohammed Farvardin. 2020. The effect of input spacing on EFL learners' vocabulary knowledge. *Language Literacy: Journal of Linguistics, Literature and Language Teaching* 4: 255–62. [CrossRef]
Sasao, Yosue. 2020. Measuring the ability to learn words. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. Abingdon: Routledge, pp. 419–32.
Schreuder, Rob, and Bert Weltens, eds. 1993. *The Bilingual Lexicon*. Amsterdam: Benjamins.
Small, Harry. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24: 265–69. [CrossRef]
Spätgens, Tessa, and Rob Schoonen. 2020. The structure of developing semantic networks: Evidence from single and multiple nominal word associations in young monolingual and bilingual readers. *Applied Psycholinguistics* 41: 1141–69. [CrossRef]
Sulpizio, Simone, Nicola Del Maschio, Davide Fedeli, and Jubin Abutalebi. 2020. Bilingual language processing: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews* 108: 834–53. [CrossRef] [PubMed]
Tai, Tzu-Yu, Howard Hao-Jan Chen, and Graeme Todd. 2020. The impact of a virtual reality app on adolescent EFL learners' vocabulary learning. *Computer Assisted Language Learning* 35: 892–917. [CrossRef]
Talib, Pawan. 2020. An introduction of corpora for ESL learners: A data driven learning about academic vocabulary. *Zanco Journal of Humanity Sciences* 24: 243–53.
Thompson, Christopher, and Sam von Gillern. 2020. Video-game based instruction for vocabulary acquisition with English language learners: A Bayesian meta-analysis. *Educational Research Review* 30: 100332. [CrossRef]
Wesche, Marjorie, and T. Sima Paribakht. 1996. Assessing vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review* 53: 13–40. [CrossRef]
Zeller, Jan. 2020. Code-switching does not equal code-switching. An event-related potentials study on switching From L2 German to L1 Russian at prepositions and nouns. *Frontiers in Psychology* 11: 1387. [CrossRef] [PubMed]
Zhang, Penchong, and Susan Graham. 2020. Vocabulary learning through listening: Comparing L2 explanations, teacher codeswitching, contrastive focus-on-form and incidental learning. *Language Teaching Research* 24: 765–84. [CrossRef]
Zhang, Yuanyue, Yao Lu, Lijuan Liang, and Baiguo Chen. 2020. The effect of semantic similarity on learning ambiguous words in a second language: An event-related potential study. *Frontiers in Psychology* 11: 1633. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

The Impact of Lexical Bundle Length on L2 Oral Proficiency

Dan Hougham [1,*], Jon Clenton [2], Takumi Uchihara [3] and George Higginbotham [4]

[1] Institute for Foreign Language Research and Education, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Hiroshima, Japan
[2] Graduate School of Humanities and Social Sciences, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Hiroshima, Japan; jclenton@hiroshima-u.ac.jp
[3] Graduate School of International Cultural Studies (GSICS), Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576, Miyagi, Japan; takumi.uchihara.a2@tohoku.ac.jp
[4] Department of Social System Design, Eikei University of Hiroshima, 1-5, Nobori-cho, Naka-ku, Hiroshima 730-0016, Hiroshima, Japan; george@eikei.ac.jp
* Correspondence: hougham@hiroshima-u.ac.jp

Abstract: Lexical bundles (LBs) are crucial in L2 oral proficiency, yet their complexity in terms of length is under-researched. This study therefore examines the relationship between longer and shorter LBs and oral proficiency among 150 L2 learners of varying proficiency levels at a UK university. Through the analysis of oral presentation data (scores ranging from intermediate to advanced) and employing a combined text-internal and text-external approach (two- to five-word bundles), this study advances an innovative text-internal LB refinement procedure, thus isolating the unique contribution of LB length. Robust regression, dominance analysis, and random forest statistical techniques reveal the predictive power of bigram mutual information (MI) and longer three-to-five-word sequences on higher proficiency scores. Our results show that learners using higher MI score bigrams tend to perform better in their presentations, with a strong positive impact on scores (b = 14.38, 95% CI [8.01, 20.76], t = 4.42; dominance weight = 58.63%). Additionally, the use of longer three-to-five-word phrases also contributes to better performance, though to a lesser extent (dominance weight = 18.80%). These findings highlight the pedagogical potential of a nuanced approach to the strategic deployment of LBs, particularly bigram MI, to foster oral proficiency. Suggestions for future LB proficiency research are discussed in relation to L2 speech production models.

Keywords: oral proficiency; multiword sequences; lexical bundles; learner corpus research; dominance analysis; random forests analysis; multiple regression analysis

1. Introduction

In recent years, there has been a growing interest in studying multiword sequences (MWSs) in second language (L2) oral proficiency. Foundational research by Pawley and Syder (1983) emphasized the critical role of lexical phrases, or "chunks", in achieving native-like fluency. They argued that a large part of native speakers' fluency stems from their use of these prefabricated chunks, which reduce the cognitive load during speech production by allowing speakers to retrieve whole phrases from memory rather than constructing sentences word-by-word. This process, known as chunking, enables more fluent and efficient language use and is crucial for effective communication.

Building on these insights, subsequent research has underscored the significance of MWSs in L2 learning and speaking proficiency (Schmidt 1992; Nation 2013). Lexical bundles (LBs), defined as frequent, contiguous MWSs, have gained particular attention for their role in speech fluency (McGuire and Larson-Hall 2021; Tavakoli and Uchihara 2020) and oral proficiency (Garner and Crossley 2018; Kyle and Crossley 2015; Zhang et al. 2021). The findings from these studies suggest that incorporating MWS-focused activities in language teaching can significantly enhance students' speaking skills. Despite this, previous studies

have predominantly focused on shorter LBs (typically two- and three-word sequences), thus potentially limiting our understanding of learners' phraseological knowledge and its impact on oral proficiency. Exploring this gap by examining both longer (e.g., *as you can see*) and shorter (e.g., *there are*) LBs is critical, as longer LBs could offer enhanced proficiency and show processing efficiencies that are essential for fluent speech production. Responding to calls from recent literature (e.g., Hougham et al. 2024; Tavakoli and Uchihara 2020), our study addresses this gap by investigating the predictive power of both shorter and longer LB usage on oral presentation scores among learners with varying proficiency levels. We aim to isolate the unique contributions of longer LBs using an innovative text-internal and text-external approach, thus providing a more comprehensive understanding of how LB length influences L2 oral proficiency. Addressing this gap is crucial for advancing our theoretical understanding of phraseological knowledge in L2 learners and for developing practical teaching strategies that leverage the benefits of various types of LBs to enhance oral proficiency.

2. Literature Review

2.1. Multiword Sequences and Lexical Bundles

Multiword sequences include various combinations of words that commonly co-occur, such as idioms (e.g., *hit the nail on the head*), collocations (e.g., *heavy rain*), phrasal verbs (e.g., *look up*), proverbs (e.g., *make hay while the sun shines*) and LBs (e.g., *in terms of*) (Garner and Crossley 2018). These sequences can be identified through either a phraseological approach or a frequency-based approach (Granger and Paquot 2008; Nesselhauf 2004). The phraseological approach categorizes MWSs using linguistic criteria (Cowie 1981), which are often based on the intuitions of first language (L1) speakers. Previous studies employing this approach (e.g., Boers et al. 2006) frequently rely on human raters to assess the formulaic nature of MWSs, but this can result in low inter-rater reliability. For example, Boers et al. (2006) reported a reliability coefficient of less than 0.60, which is significantly lower than the median inter-rater reliability of 0.92 reported in SLA research by Plonsky and Derrick (2016).

On the other hand, the frequency-based approach, developed by Sinclair (1991) and Nesselhauf (2004), uses corpus-based automated extraction techniques to identify frequent word combinations based on quantitative criteria such as frequency and range. These are often termed lexical bundles or n-grams and can include both structurally complete (e.g., *at the end of the day*) and incomplete sequences (e.g., *in the middle of*), regardless of their idiomaticity or structural status (Biber et al. 1999). The objective nature of the frequency-based approach has made it popular in learner corpus research (Paquot and Granger 2012). For instance, Ebeling and Hasselgård (2015) emphasized the benefits of this approach, such as accessing a large quantity of data for quantitative analysis.

In our study, we use the term LB to refer to sequences identified using automatic extraction software based on criteria like minimum frequency, range, and mutual information (MI) scores, which measure the strength of association between word pairs. Our aim is to distinguish LBs from other MWSs, even though there may be some conceptual overlap (e.g., high MI score bigrams like *global warming* can also be categorized as collocations).

In learner corpus research, LB production can be analyzed using either text-internal or text-external approaches. The text-internal approach (e.g., Biber and Gray 2013) focuses on data within learner corpora. In contrast, the text-external approach (e.g., Tavakoli and Uchihara 2020) examines LBs in terms of criteria like the frequency of co-occurrence in external reference corpora consisting of L1 speaker language. Both approaches have their limitations such as the arbitrary frequency cutoffs in the text-internal approach (Myles and Cordier 2017) and the challenges in measuring longer bundles with text-external analysis tools (e.g., some software programs such as the Tool for the Automatic Analysis of Lexical Sophistication: TAALES (Kyle et al. 2018) can only analyze bigrams and trigrams). Additionally, text-external techniques may not ensure that frequently occurring word combinations in external corpora are psycholinguistically relevant to the learners being

studied (Ellis et al. 2009; Myles and Cordier 2017). Our study employs both approaches to assess the contributions of shorter and longer LBs with respect to L2 oral proficiency.

2.2. From Theory to Empirical Findings

Pawley and Syder (1983) lay foundational insights into the importance of multiword sequences in language proficiency through their exploration of lexical phrases. They propose that a speaker's command of fixed and semifixed phrases, or "chunks" of language, is crucial for achieving native-like fluency. Such lexical phrases ease the cognitive load during speech production and serve as essential building blocks for fluent and idiomatic language use. This early perspective is significant, as it suggests that the repertoire of lexical phrases is a key factor distinguishing more proficient speakers from less proficient ones, thereby enabling more efficient processing and production of language. Building upon the understanding of lexical phrases, Levelt's (1989) speech production model provides further theoretical justification for the link between LBs and proficiency. The model describes three stages in speech production: conceptualization, formulation, and articulation. During the conceptualization stage, speakers plan the content of their speech. In the formulation stage, they encode lexical and grammatical items in the mental lexicon, activating appropriate lemmas and constructing syntactic structures. Finally, the articulation stage involves implementing the phonetic plan, resulting in speech production. Originally designed for L1 speakers, Kormos (2006) adapted this model to address the specific challenges of L2 speakers, such as a smaller and less structured mental lexicon with fewer formulaic expressions. The model suggests that speakers with a larger repertoire of MWSs can retrieve these sequences as easily as single words during the formulation stage, which reduces cognitive load and provides a processing advantage. This efficiency allows L2 learners to allocate cognitive resources to other aspects of speech production, such as lexical and grammatical accuracy or complexity (Kormos 2006; Skehan 2014). Utilizing MWSs can boost oral fluency, particularly during lexical selection at the formulation stage (Kormos 2006; Levelt 1992). In contrast, speakers with a limited MWS repertoire may struggle, as they expend more cognitive resources when retrieving individual lexical items during the formulation stage. Alongside these frameworks, usage-based theory, as articulated by Ellis (2002) and supported by Bybee (2006), further complements our understanding of how language proficiency develops from the experience and frequency of usage. Ellis (2002) introduces the notion that linguistic competence is significantly influenced by the accumulation of exposure to MWSs, which reinforces their retrieval and production. This perspective aligns with the proceduralization of language use, where proficiency emerges from the frequent and familiar use of language structures. Bybee suggests that repeated exposure to MWSs not only facilitates their entrenchment in the learner's mental lexicon but also highlights the importance of frequency and familiarity for L2 learning. Together, these theoretical perspectives—Pawley and Syder's lexical phrases theory, Levelt's speech production model, and the usage-based theories proposed by Ellis and Bybee—form a multifaceted view of the mechanisms underlying L2 oral proficiency. The current study aims to bridge these theoretical insights with empirical evidence, exploring how the length, frequency, and association strength of LBs contribute to oral proficiency. We seek to illuminate the relationship between LB use and language proficiency, grounded in a broad spectrum of linguistic theory.

To date, most learner corpus-based studies have focused on LBs used in writing rather than speaking (e.g., Appel and Wood 2016; Siyanova-Chanturia and Spina 2020; Staples et al. 2013). This emphasis can be attributed to two main reasons. First, natural speech within L2 classes/tests is often not recorded, whereas L2 students regularly submit written work, thus making written corpora more readily available for analysis. Second, spoken corpora require more time to analyze, as there are additional steps in processing. Specifically, the speech first needs to be transcribed into a text form, and it may need to be divided into speech units for ease of analysis, thus making the analysis of spoken language more labor-intensive and time-consuming than written texts. Given these constraints,

most research has naturally gravitated towards written corpora. However, the findings from studies focusing on written language—highlighting the complexity of the MWS–L2 proficiency link and trends regarding the use and proficiency-related evolution of LBs—underscore a rich area for exploration within spoken language as well. Specifically, the research exploring written LBs has revealed general trends, including the following: (a) the quantity of LBs produced by learners decreases as proficiency increases (Appel and Wood 2016; Staples et al. 2013) or with time spent in an English-speaking country (Groom 2009). Less proficient learners overuse bundle tokens and under-use bundle types, which is a trend resembling the use of "phraseological teddy bears" (i.e., overusing high-frequency phrases with which one feels comfortable, as demonstrated in Hasselgård 2019). Research has shown that proficient learners have a firmer and more creative command of lower-frequency bundles whose constituent words are nonassociated (Siyanova-Chanturia and Spina 2020); and (b) less proficient learners rely more on LBs copied from writing prompts or source texts due to their relatively limited lexical repertoires (e.g., Appel and Wood 2016; Staples et al. 2013). Overall, these studies suggest that much remains to be understood about the relationship between LB use and general writing proficiency. While most existing learner corpus-based studies have focused on the use of LBs in written contexts, the insights they provide into the relationship between MWS–L2 proficiency potentially apply to the spoken domain as well. The understanding that learners' reliance on LBs decreases with proficiency or extended exposure to an English-speaking environment, and that proficiency dictates a learner's command over lower-frequency bundles, offers a compelling framework to consider for spoken data. The current study seeks to bridge this gap by exploring how these patterns manifest in speech proficiency, which has been less frequently examined. There is potential that the dynamics of LB use in writing, as detailed in these studies, could find parallels in oral production, thereby providing richer insights into speech proficiency and its intricacies.

Several recent studies have examined the link between LB use and general speaking proficiency from a text-external perspective, focusing on the extent to which L2 learners use L1 target-like two- and three-word (bi- and trigram) measures in terms of quantitative indices (frequency, proportion, and association) (e.g., Garner and Crossley 2018; Kyle and Crossley 2015; Zhang et al. 2021). The findings from these studies differ from the typical trend of writing-centered studies (which proposed that less-proficient L2 writers (over)use a greater number of high-frequency bundles). Kyle and Crossley (2015) found significant correlations between (human-rated) oral proficiency scores and several n-gram scores, of which the strongest predictor of speaking proficiency came from high-frequency trigrams, thus suggesting that more skillful L2 speakers use a larger number of highly frequent trigrams. Garner and Crossley (2018) found that beginning-level L2 learners showed the greatest increase in oral production of high-frequency bigrams over the course of their four-month longitudinal study. Zhang et al. (2021) reported that several n-gram measures (e.g., bigram proportion and association: MI and t scores) significantly correlated with (human-rated) oral proficiency scores on story retelling and monologic tasks. Such studies highlight the important role of proficiency in the development of LB use but suggest that further research is required to bring clarity to this research area.

Relatively few studies (e.g., Biber and Gray 2013; De Cock 2004) have examined LB use in spoken corpora from a text-internal perspective. De Cock (2004) examined two- to six-word bundle use among advanced EFL learners compared to L1 speakers. She found that learners' preferred bundles were less interactional and included relatively few vagueness markers (e.g., *or something*, *kind of*) compared to L1 speakers. Biber and Gray's (2013) study of spoken and written responses to the TOEFL iBT showed a slightly more complex pattern than other n-gram studies. They reported that intermediate-level participants produced a greater number of bundles (four-word units) than their lower- and higher-proficiency groups, thus suggesting a general developmental progression in which lower-level participants use a smaller number of bundles, intermediate-level participants overuse a larger number of bundles, and high-scoring participants show greater control

and creativity in using the bundles they have acquired (p. 37). To summarize, these studies suggest a complex perspective and a need for further research on how language use varies among individuals from diverse backgrounds.

While such studies have offered insights into the patterns of LB use in spoken corpora, a broader context emerges when we consider the significant relationship between MWS use and speech fluency, and it is clear across various teaching contexts. Studies that show significant and positive relationships between MWS use and speech fluency (including L2 proficiency in the broader sense) come from a range of different teaching contexts (e.g., Boers et al. 2006; Hougham et al. 2024; McGuire and Larson-Hall 2017, 2021; Stengers et al. 2011; Suzuki et al. 2022; Tavakoli 2011; Tavakoli and Uchihara 2020; Uchihara et al. 2021; Wood 2009, 2010). Boers et al. (2006) and Stengers et al. (2011) found strong links between the number of MWSs used (in story retelling tasks) and (perceived) oral ability scores in a Belgian EFL context. Wood (2009) examined the effect of MWS-focused teaching on MWS use and oral fluency in a case study ($N = 1$) in a Canadian ESL context. Wood found that MWS-focused instruction can lead to increased MWS use and increased spoken fluency over a short period (six weeks). Wood (2010) also found similar results with a slightly larger sample size ($N = 11$) in a similar context over a longer period (six months). Tavakoli (2011) compared the pausing patterns of L1 versus L2 speakers' performance in a UK university context. She found that L2 learners rarely paused in the middle of multi-word units, thus providing further corroborating evidence that lexical chunks facilitate fluency. Similarly, Uchihara et al. (2021) found that speakers who provided more low-frequency MWS (collocational type) responses to a word association task (Lex30) spoke more rapidly with fewer silent pauses. McGuire and Larson-Hall (2017) replicated Wood's (2009) study in an American ESL study abroad context. They reported a moderately strong relationship between all participants' MWS use and fluency measures. Tavakoli and Uchihara's (2020) study, reporting the link between two- and three-word LBs and one objective measure from each aspect of utterance fluency (speed, breakdown, and repair) across assessed proficiency levels in a UK university context, represents the first systematic study of its kind. Tavakoli and Uchihara reported that greater LB use (a larger proportion of frequent LBs and more frequent LBs) was positively and significantly related to higher speaking ability scores and with some fluency aspects (faster articulation rate and fewer pauses within clauses). Suzuki et al.'s (2022) task repetition intervention study examined the use of single words and trigrams on speed, breakdown, and repair fluency aspects. They found that the recycling of more complex MWSs through task repetitions seemed to facilitate proceduralization (i.e., more efficient retrieval of MWSs), but they also found that such reuse had both positive and negative influences on midclause pauses specifically, as well as fewer but longer pauses within clauses, which may show that learner encoding systems were in the process of restructuring. Hougham et al. (2024) examined the relationship between both shorter (bi- and trigram) and longer (four–to-five-word) LBs and three dimensions of fluency (speed, breakdown, and repair) using both text-internal and text-external techniques. They found that using longer LBs, specifically four-to-five-word sequences of high collocational quality (those with high MI scores), significantly enhanced speech fluency by reducing the frequency of pauses and repairs. Moreover, they uncovered a correlation between frequent combinations of two words and a faster rate of speech, whereas complex combinations (those with high mutual information scores) were found to slow down speech.

2.3. Gaps and Unexplored Areas in Previous Research

While the studies reviewed here support the hypothesis of a positive relationship between MWS use and proficiency, they might be limited in at least six important ways. Foremost, studies focusing on relatively short (two- and/or three-word) sequences might not fully capture learners' actual phraseological knowledge and how it relates to oral fluency units. This is exemplified by multiple studies (e.g., Garner and Crossley 2018; Kyle and Crossley 2015; Kyle et al. 2018; McGuire and Larson-Hall 2021; Suzuki et al. 2022; Tavakoli and Uchihara 2020; Zhang et al. 2021). For example, using longer LBs might be

more beneficial for improving aspects of oral fluency and increasing high-stakes assessment scores. Emphasizing the potential importance of longer LBs, Tremblay et al. (2011) demonstrated that longer (four- and five-word) LBs offer online processing advantages over non-LBs in receptive tasks. Hougham et al. (2024) also showed that longer LBs of high collocational quality enhanced various aspects of fluency. Despite this understanding, the effects of longer linguistic units on achieving fluency in speech production have not been fully explored. Given what we know from Pawley and Syder's insights into the significance of lexical phrases, the comprehensive framework provided by Levelt's speech production model, as well as the principles of usage-based theory by researchers such as Bybee and Ellis, it is reasonable to hypothesize that employing longer LBs can lead to enhanced processing efficiency. Second, many previous studies have had methodological or contextual limitations, such as measuring MWSs subjectively using a criteria checklist and L1 speaker intuition (e.g., McGuire and Larson-Hall 2017; Wood 2009). Third, most previous research is restricted to investigating the MWS proficiency link with a learner-external approach (i.e., examining learners' use of selected sequences that are thought to be formulaic in L1 speaker English and identified in advance as formulaic or quantifying a text's formulaicity by checking the frequency of all of its constituent word sequences against an external reference corpus) (e.g., Garner and Crossley 2018; Tavakoli and Uchihara 2020; Zhang et al. 2021). Only one study by Hougham et al. (2024) has systematically attempted to employ both text-external and text-internal methods to analyze learner-produced LBs in relation to aspects of oral fluency, but their study did not examine oral proficiency scores. Fourth, some previous studies have suffered from small sample sizes (e.g., $N = 19$ in McGuire and Larson-Hall 2017; $N = 1$ in Wood 2009; $N = 11$ in Wood 2010). Fifth, the LBs examined in studies based on learner corpora frequently have varying lengths, typically ranging from two to six words. Additionally, the criteria for extracting these LBs, such as frequency and dispersion, differ significantly across different studies. Therefore, it is important to view the aforementioned findings and general patterns as hypotheses that require further testing using alternative corpus data within diverse contexts across various proficiency levels.

3. The Current Study

Informed by prior theoretical insights (e.g., Levelt 1989; Pawley and Syder 1983) and responding to the call for a more comprehensive approach to MWSs in EFL research (Hougham et al. 2024; Tavakoli and Uchihara 2020), our study explores the under-investigated area of longer LB usage, hypothesizing that these can offer significant processing advantages for L2 speakers, which is a notion previously suggested but not empirically tested across a range of proficiency levels. Our research question is designed to directly address the identified need for more comprehensive analyses of LB usage in relation to proficiency scores. This is done by building on and extending the work of Tavakoli and Uchihara (2020), as well as Hougham et al. (2024). We investigate the relationship between the use of shorter (bi- and trigrams) and longer (three-to-five-word) LBs and oral proficiency scores. By comparing and contrasting findings across different learner populations and proficiency scores, the current study seeks to contribute to a more detailed understanding of how LBs function in L2 speech production models and inform future LB proficiency research directions. By examining a broader dataset (150 L2 learners at varying proficiency levels from a UK university's presessional course) and using an innovative text-internal LB refinement procedure that identifies more structurally complete and useful LBs, the current study allows us to explore the relationship between LB usage and oral proficiency, aiming to provide a more comprehensive understanding of the relationship. Additionally, the current study seeks to explore the extent to which both shorter and longer LBs can predict speaking proficiency scores. By employing robust regression, dominance analysis, and random forest techniques, we not only aim to validate previous findings but also uncover new patterns, potentially leading to more effective pedagogical strategies for enhancing oral proficiency in diverse L2 learning contexts.

The current study addresses the following research question: To what extent do longer lexical bundles (three-to-five-word units) predict oral proficiency scores compared to shorter lexical bundles (bi- and trigrams) in L2 learners?

Based on the findings in Tavakoli and Uchihara (2020) and the theoretical frameworks of speech production (e.g., Kormos 2006), the current study hypothesizes that longer LBs will significantly contribute to higher oral proficiency scores, potentially more so than shorter LBs, due to their ability to enhance processing efficiency and fluency.

4. Materials and Methods

4.1. Participants

The participants were 150 language learners taking a presessional course at a UK University. They were from 20 different L1 backgrounds (outlined in Table 1), with most of the participants being L1 Chinese (n = 101), L1 Saudi Arabian (n = 9), or L1 Turkish (n = 9). Participants' raw presentation scores (described in the next section) were first converted into IELTS bands ranging from 6.5 to 7.5. Next, they were categorized into three groups depending on these bands. From a larger pool, we randomly selected 50 individuals for each level, ensuring balanced representation across the three bands. There were 50 participants at IELTS level 6.5, 50 at IELTS level 7, and 50 at IELTS level 7.5 (see Table 2).

Table 1. Participants according to L1 backgrounds (N = 150).

Nationality	*n*	Nationality	*n*	Nationality	*n*
Chinese	101	Indonesian	2	Taiwanese	1
Saudi Arabian	9	Bangladeshi	1	Chilean	1
Turkish	9	German	1	Egyptian	1
Japanese	7	Bahraini	1	Georgian	1
Thai	6	Kuwaiti	1	Pakistani	1
Indian	2	Ghanian	1	Russian	1
Colombian	2	Italian	1		

Table 2. Participants' IELTS-rated speaking proficiency levels.

IELTS speaking level	6.5	7	7.5	Total
n	50	50	50	150

4.2. Oral Presentation Tasks and Proficiency Scores

All participants completed oral presentations comprising a 7 min presentation (monologue) in small groups with the aid of PowerPoint slides as part of the presessional course. It is important to note that these presentations were not conducted specifically for the purposes of the current study but were completed as a requirement of the presessional course at Queen Mary University of London. Their outputs were video recorded online and assessed by teachers using grading descriptors (in equal measure): presentation content, presentation structure, seminar leadership, language fluency, and language accuracy (see the full descriptors in Appendix A). The oral presentation tasks were designed to cover a wide range of subtopics under the overarching theme of globalization, which had been the focal point of the participants' 5-week academic English course. While globalization served as the central theme, participants were encouraged to explore this broad topic through various lenses, ranging from economic and legal perspectives (e.g., how globalization affects development and international trade law) to more specialized areas (e.g., acoustic telemetry in fisheries). This diverse range of subtopics allowed participants to delve into areas aligned with their academic interests and expertise. The students were from a wide range of disciplines: humanities, law, science, technology, engineering, and mathematics majors. From the video recordings, we took 3 min speech samples starting at the 30 s timestamp. As the 3 min sections of speech analyzed came from the "presentation" part of

their seminar, we are specifically analyzing monologic oral presentations rather than freer dialogic speech. These samples were transcribed using Sonix.ai web-based software, and the transcriptions were checked for accuracy by a research assistant and double-checked by the first researcher. We used the transcripts in the lexical analyses (described in detail below), and we used the raw scores and the banded IELTS scores as the measures of oral proficiency in the current study.

4.3. Measuring Lexical Bundles: A Two-Pronged Approach

For the current study, we adopt a frequency-based approach using both text-internal and text-external techniques to isolate the unique contribution of shorter versus longer LBs.

4.3.1. Text-External Lexical Bundle Analysis and Measures

Following previous studies (Garner and Crossley 2018; Hougham et al. 2024; Tavakoli and Uchihara 2020), we used three n-gram indices (proportion, frequency, and association) to objectively measure the use of shorter LBs, specifically two- and three-word contiguous sequences (i.e., bi- and trigram tokens) in our learner corpus. TAALES version 2.0 was used to calculate three kinds of n-gram scores, producing six score indices (two proportion, two frequency, and two association indices). As our external reference corpus, we chose the spoken subsection of the Corpus of Contemporary American English (COCA Davies 2009), which comprises 79 million words from transcriptions of a wide range of TV and radio programs. Our choice of this spoken corpora was in alignment with research findings showing a gap in L2 learners' spoken and written vocabulary sizes (Uchihara and Harada 2018) and differences in lexical profiles between spoken and written modes (Dang et al. 2017). We maintained consistency between the modality in which L2 words were elicited and the modality of the reference corpus based on the practices used in previous studies (e.g., Uchihara and Clenton 2020; Uchihara et al. 2021).

In our current study, we use proportion score indices to measure the occurrence of bi- and trigrams in our learner speech sample data. These bi- and trigrams are also among the 30,000 most frequent ones in the external reference corpus (COCA). Higher proportion scores show that participants in the sample produced a higher percentage of high-frequency, target-like bi- and trigrams. Higher frequency scores show that participants in our sample produced a larger number of high-frequency target-like bi- and trigrams. Logarithmic bi- and trigram scores, instead of raw frequency scores, were used to control for Zipfian effects common in word frequency lists (Kyle and Crossley 2015; Tavakoli and Uchihara 2020). Association score indices measure the association strength between individual words within bigrams and trigrams. Of the five association measures available in TAALES, the one association measure we used was mutual information score.[1] MI score measures the strength of association between two words. MI scores show the strength of word associations, with higher scores suggesting stronger associations. However, MI also focuses on word pairs that are not commonly found together (Schmitt 2010, p. 130). Before n-gram analysis using TAALES, all the transcripts were cleaned by correcting any misspellings and mispronunciations and removing any markings of filled pausing (i.e., *ums*, *uhs*, etc.). The resulting transcripts ranged between 216 and 436 words (M = 310.78, SD = 47.32).

4.3.2. Text-Internal Lexical Bundle Identification and Refinement Procedures

We adopted a text-internal approach to isolate and measure the unique contribution of longer LBs to proficiency. As a first step, we conducted frequency analyses using AntConc (Anthony 2022) to generate lists of the most frequently used four-word LBs in the learner corpus. The frequency and dispersion thresholds used to identify lexical bundles vary from study to study. Figures used for "frequency cut offs are somewhat arbitrary" (Hyland 2008, p. 8) depending on both the size and specificity of the corpus. For relatively small spoken corpora like the one in this study, a raw cutoff frequency has often been used, ranging from two to ten occurrences (e.g., Altenberg 1998; Biber and Barbieri 2007; De Cock 1998). Given the small size of the spoken corpus in the current study

(46,617 words), for four-word combinations to qualify as lexical bundles, we used a cutoff point of three or more occurrences in at least three texts, following Biber and Barbieri (2007). These minimum figures help to ensure that the identified bundles are not idiosyncrasies confined to occurrences produced by an individual speaker. This resulted in 447 instances of four-word LBs that met these criteria. To deal with the issues of overlap and structural incompleteness among these instances, we used a refinement procedure developed by previous researchers (Wood and Appel 2014) aiming to identify more structurally complete and useful LBs. We split each four-word sequence (e.g., *I will give you*) into two constituent three-word clusters (e.g., *I will give, will give you*). The frequency of the two three-word clusters in the corpora were identified and compared. If the frequency of one three-word cluster was at least double the frequency of the other, the more frequent cluster was classified as the root structure and the fourth word was considered as a word that commonly occurred with that structure and was put in parentheses. For example, *I will give* occurred over two times (freq = 41) than *will give you* (freq = 20) in the current data set. Therefore, the final resultant structure is in this case: *I will give (you).* Another example from the current data set is the sequence *the first part is*. Since *the first part* (freq = 41) occurred over two times more frequently than *first part is* (freq = 17), the final resultant structure is *the first part (is)*. The refinement process produced a list of 119 multiword structures, primarily comprising core three-word phrases and four-word structures, with a smaller number of longer five-word structures included. Table 3 shows the number of three-word (55), four-word (58) and five-word structures (6) identified. The examination of the extensive list provided in Appendix B shows that many of the resulting three-word structures are self-contained units in terms of semantics or structure. For instance, *with the development of* forms a complete unit. This observation suggests that the refinement procedure successfully pinpointed additional core structures.

Table 3. Numbers and examples of structures identified at different lengths.

Lexical bundle length	three-word units	four-word units	five-word units	Total
Number of units	55	58	6	119
Examples	*I'm going to* *the first part* *I will give*	*I would like to* *at the end of* *at the same time*	*if you have any questions* *the presentation will be brief* *in a socially responsible way*	

Note: Contracted forms (e.g., *I'm*) were counted as a single word; for example, *I'm going to* was considered to be a trigram rather than a quadgram.

It is crucial to differentiate the newly identified structures from traditional LBs (and other multiword structures) when describing them, as this refinement procedure goes beyond the usual LB approach. Recall that the traditional LB approach strictly identifies multiword sequences within two parameters: frequency and range. For this reason, when making modifications or refinements to traditional LB methods, previous researchers (e.g., Simpson-Vlach and Ellis 2010; Wood and Appel 2014) have used different terminology to refer to such refined or modified LBs. Simpson-Vlach and Ellis (2010), for instance, started with an LB approach and then applied additional criteria (e.g., human ratings of formula teaching worth combined with MI score as a measure of collocation strength), thus aiming to identify more useful multiword units for teaching purposes. Simpson-Vlach and Ellis (2010) adopted the general term "formulaic language" to describe the word combinations identified in their study. Another example is Wood and Appel (2014), who developed the LB refinement procedure used in the current study. Wood and Appel adopted the general term "multiword constructions" to refer to the refined LBs identified in their study. "Formulaic language" and "multiword constructions" have also been used as umbrella terms in other studies such as Liu (2012). Because of the fuzzy nature of boundaries between many types of semifixed multiword combinations, and because each study's identification and refinement methods are different, there is no consensus in the literature as to which terms apply in all cases. Although it is challenging to pin down a consistent definition of

MWSs across all studies, most researchers agree that it is important to distinguish between different types of MWSs where possible and to report clearly how MWSs are identified in each study to facilitate comparisons of research findings across studies. In the current study, to avoid terminological confusion, we use the term "refined LBs" to refer to the refined LB list (see Appendix B) produced by the above-described refinement procedure.

4.4. Scoring the Use of Text-Internal Bundles

It is important to quantify usage of the list of refined LBs so that we can run various quantitative analyses (e.g., multiple regression) and compare the relationships between different types of LBs (unrefined text-external vs. refined text-internal; shorter vs. longer) with oral presentation scores. To do so, we awarded one point for each identified refined LB used by each participant. We tallied up the total number of points for each participant, arriving at a three-, four-, and five-word usage score for each participant. As for text-internal MI scores, we extracted MI scores for sequences of various lengths using the Collocate 2.0 software program (Barlow 2015). MI scores for three-to-five-word sequences have been used in several studies (e.g., Ellis et al. 2008; Simpson-Vlach and Ellis 2010), as they appear to offer a reliable indication of phrasal coherence. For each refined bundle used by each participant, we awarded the corresponding MI score. We then tallied all MI scores and gave each participant a total MI score.

4.5. Statistical Analyses

We analyzed six text-external n-gram (two- and three-word) measures and two text-internal refined n-gram (three-, four-, and five-word) measures. To examine our research question, that the use of LBs of varying length can predict oral presentation raw scores, we selected regression analysis. Initially, we examined the assumptions underlying regression models. The presentation raw scores variable yielded a Shapiro–Wilk p-value of 0.04, thus showing a significant deviation from a normal distribution. To identify influential outliers within the data, Cook's distance was employed with a conservative threshold of $4/n$, thus facilitating the assessment of the impact of individual data points on the regression model. Several data points emerged as influential outliers, surpassing the Cook's distance threshold. Their presence implies a potential influence on the estimated regression coefficients (see Supplementary Materials for detailed results of these checks). Subsequently, guided by Larson-Hall (2015, p. 264), we conducted a robust regression using MM estimation with the "rlm" function in the MASS package in R (R Development Core Team 2019). We chose the "rlm" function because it aptly accommodates data sets with non-normal distributions and outliers.

In an effort to make the results more robust, we decided it was relevant to conduct additional analyses on the relative importance of each predictor variable in explaining the variance in the model. We pursued this inquiry through a dominance analysis (DA) using the "calc.relimp" function from the "relaimpo" package in R (Grömping 2006). DA can effectively address correlations among predictor variables and can help in better understanding the unique contribution of each PV to the criterion variable in multiple regression analysis as opposed to relying solely on possibly misleading standardized beta coefficients (Mizumoto 2022). DA facilitates comprehension by computing dominance weights for each predictor, which show the mean impact of a variable on the predictability of all potential subsets of predictors, consequently presenting a thorough picture of the influence of each predictor on the outcome.

To achieve a more precise estimation of the importance of each variable in the multiple regression model, it is important to conduct dominance analysis in combination with random forests analysis (Mizumoto 2022). The random forests approach is a nonparametric machine learning model, meaning it can offer more precise outcomes when multiple regression assumptions are violated (Liakhovitski et al. 2010). Using random forests allows researchers to acquire a nuanced perspective on variable importance. Hence, following the guidelines by Mizumoto (2022), we integrated the random forests analysis using

the Boruta package in R (Kursa and Rudnicki 2010). The Boruta algorithm, specifically designed for feature ranking based on random forests, runs the random forests multiple times. It labels features (or predictors) as "confirmed", "rejected", or "tentative" based on their significance compared to randomized shadow features. "Confirmed" predictors are deemed significant, "rejected" ones are considered unimportant, and "tentative" labels are reserved for predictors whose importance remains uncertain. We show these using boxplots in the figures presented in the following section.

Integrating Boruta with robust regression and DA allows for a comprehensive analysis, where each method compensates for the limitations of the others. Robust regression ensures that our model is not unduly influenced by outliers, providing reliable coefficient estimates even when the data distribution is non-normal. DA helps clarify the relative importance of each predictor by showing their unique contributions to the model while addressing correlations among predictors. Meanwhile, Boruta offers a rigorous feature selection process that ranks predictors based on their importance and independently of the distributional assumptions. By using Boruta, we can validate which predictors are truly significant, providing a layer of verification to the results obtained from robust regression and DA. This combination of methods allows for triangulation, where the results from one method can support and validate the findings of the others, thus possibly enhancing the overall reliability and robustness of our conclusions. In what follows, we present the descriptive statistics first, we then report the robust regression model in combination with DA and random forests to address our research question.

5. Results

Table 4 shows the descriptive statistics for the presentation raw scores and different LB measures used in this study. The table includes the mean (*M*), standard deviation (*SD*), median, minimum, and maximum values for each measure. These statistics provide an overview of the distribution and central tendency of the presentation scores and n-gram measures, highlighting the variability and range of the data.

Table 4. Descriptive statistics for presentation scores and n-gram measures.

Measure	*M*	*SD*	Median	Minimum	Maximum
Presentation raw scores	62.71	5.53	63.00	53.00	78.00
Bigram log frequency	1.32	0.12	1.31	1.04	1.73
Bigram proportion	0.44	0.07	0.44	0.26	0.63
Bigram association MI	1.54	0.17	1.54	1.13	2.09
Trigram log frequency	0.68	0.13	0.67	0.41	1.07
Trigram proportion	0.12	0.04	0.12	0.04	0.26
Trigram association MI	2.53	0.25	2.54	1.91	3.21
Three-to-five-word usage	10.29	5.10	10.00	1.00	27.00
Three-to-five-word MI	132.39	73.57	117.97	19.87	399.56

Note: M = mean. *SD* = standard deviation. MI = mutual information.

A boxplot was created to visualize the distribution of presentation scores across different L1 groups (Figure 1). The boxplot provides a clear representation of the central tendency and variability within each group. Although a Kruskal–Wallis test did not find statistically significant differences in the presentation scores across L1 groups (Kruskal–Wallis chi-squared = 29.825, df = 19, *p*-value = 0.054), the boxplot offers valuable insights into the data distribution. It shows that most L1 groups have similar median scores, but there is considerable variability within some groups. The blue dots represent individual data points, highlighting the spread of scores within each group.

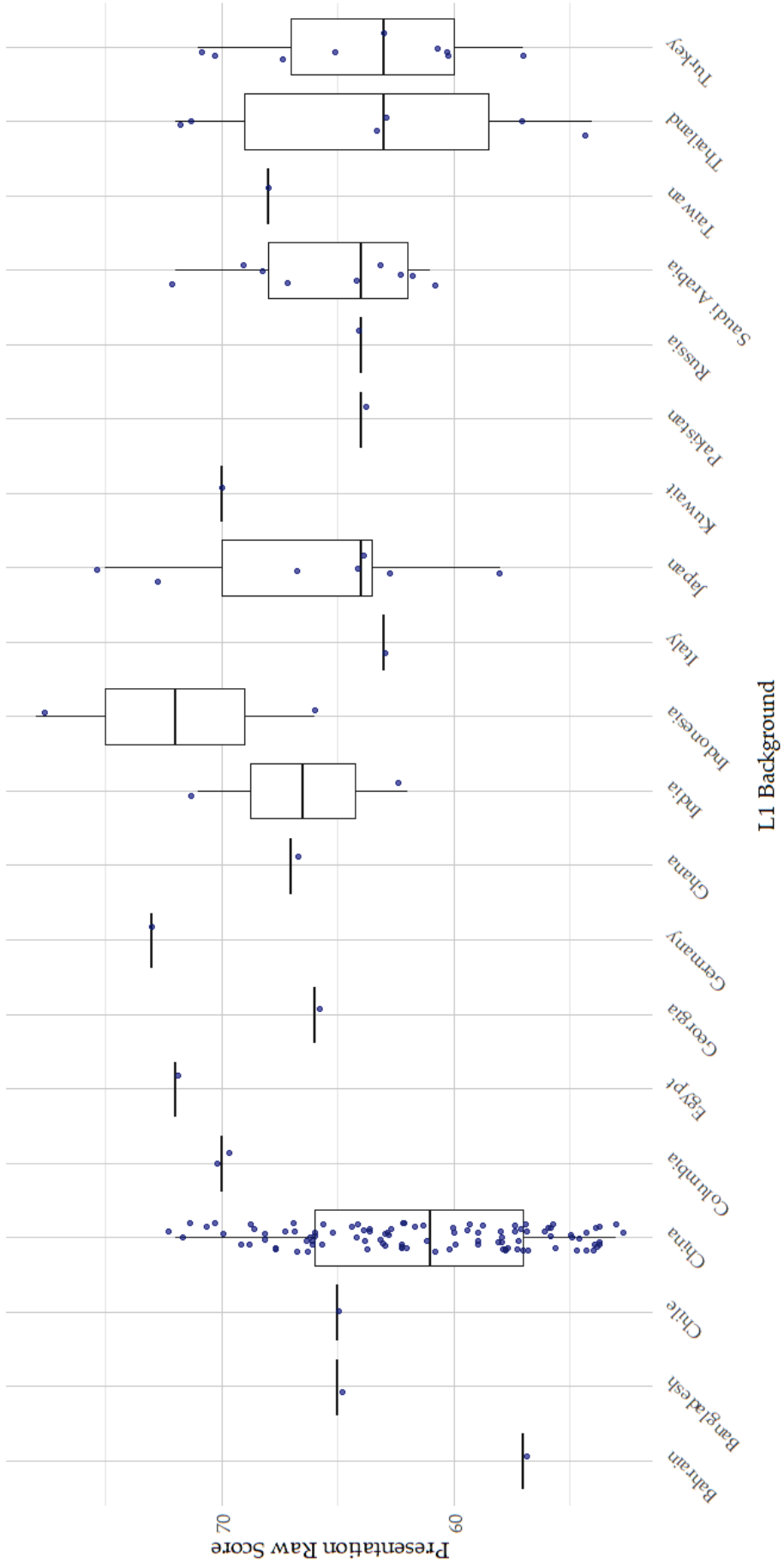

Figure 1. Presentation scores according to L1 background.

Table 5 shows the findings from the robust regression and dominance analysis with the criterion variable being presentation raw scores. The robust regression analysis unveils multiple predictors impacting the presentation raw scores. Bigram MI displayed a strong positive association (b = 14.38, 95% CI [8.01, 20.76], t = 4.42), accounting for a dominant 58.63% of the variance in presentation raw scores. This implies that learners using more bigrams with higher MI scores tend to score higher on their presentations, holding other variables constant. Examples of bigrams with high MI scores produced by high-scoring participants include *carbon dioxide, Kyoto Protocol, and global warming* (for more examples, see Appendix C). Three-to-five-word usage also showed a significant positive influence on the model (b = 0.53, 95% CI [0.06, 0.99], t = 2.22), contributing to 18.80% of the variance in the presentation raw scores. This suggests that individuals employing specific three-to-five-word phrases achieve higher presentation scores. Other predictors like trigram frequency, trigram proportion, and trigram MI indicated lesser influence, with each contributing less than 4% to the total dominance weight.

Table 5. Robust regression and dominance analysis (criterion: presentation raw scores).

Predictor	B	95% CI	SE	T	Dominance Weight (%)
Intercept	39.02	16.45, 61.58	11.51	3.39	
Bigram frequency	8.10	−4.17, 20.37	6.26	1.29	0.008 (3.41%)
Bigram proportion	−12.09	−39.57, 15.39	14.02	−0.86	0.006 (2.64%)
Bigram MI	14.38 *	8.01, 20.76	3.25	4.42	0.129 (58.63%)
Trigram frequency	−3.07	−13.10, 6.95	5.11	−0.60	0.003 (1.23%)
Trigram proportion	4.15	−58.81, 67.11	32.12	0.13	0.007 (3.23%)
Trigram MI	−2.05	−6.24, 2.14	2.14	−0.96	0.007 (3.28%)
Three-to-five-word usage	0.53*	0.06, 0.99	0.24	2.22	0.041 (18.80%)
Three-to-five-word MI	−0.02	−0.05, 0.01	0.02	−1.25	0.019 (8.78%)
Total					0.220 (100%)

Note: MI = mutual information. * Results are marked significant if the 95% confidence interval excludes zero.

In Figure 2, the dominance weights are presented in descending order starting from the predictor variable with the highest weight (bigram MI) and ending with the one with the lowest weight (trigram frequency). Figure 1 helps us visualize each predictor's relative importance, highlighting bigram MI as the most important predictor among all variables in our study.

Figure 3 shows a variable importance plot derived from random forests using the Boruta algorithm. The Boruta results confirm the importance of five attributes: Bigram MI, three-to-five-word usage, trigram proportion, three-to-five-word MI, and bigram proportion. Two attributes, trigram frequency and trigram MI, were confirmed as unimportant, while bigram frequency remained tentative. Detailed results and the R code are available in Supplementary Materials.

Overall, the Boruta analysis, along with dominance and robust regression analyses, together highlight the substantial influence of bigram MI and three-to-five-word usage on presentation scores. These consistent findings across different analytical techniques underscore the importance of specific lexical choices, especially bigram MI, in determining speaking performance, thus offering a triangulated insight into how LBs impact speaking proficiency.

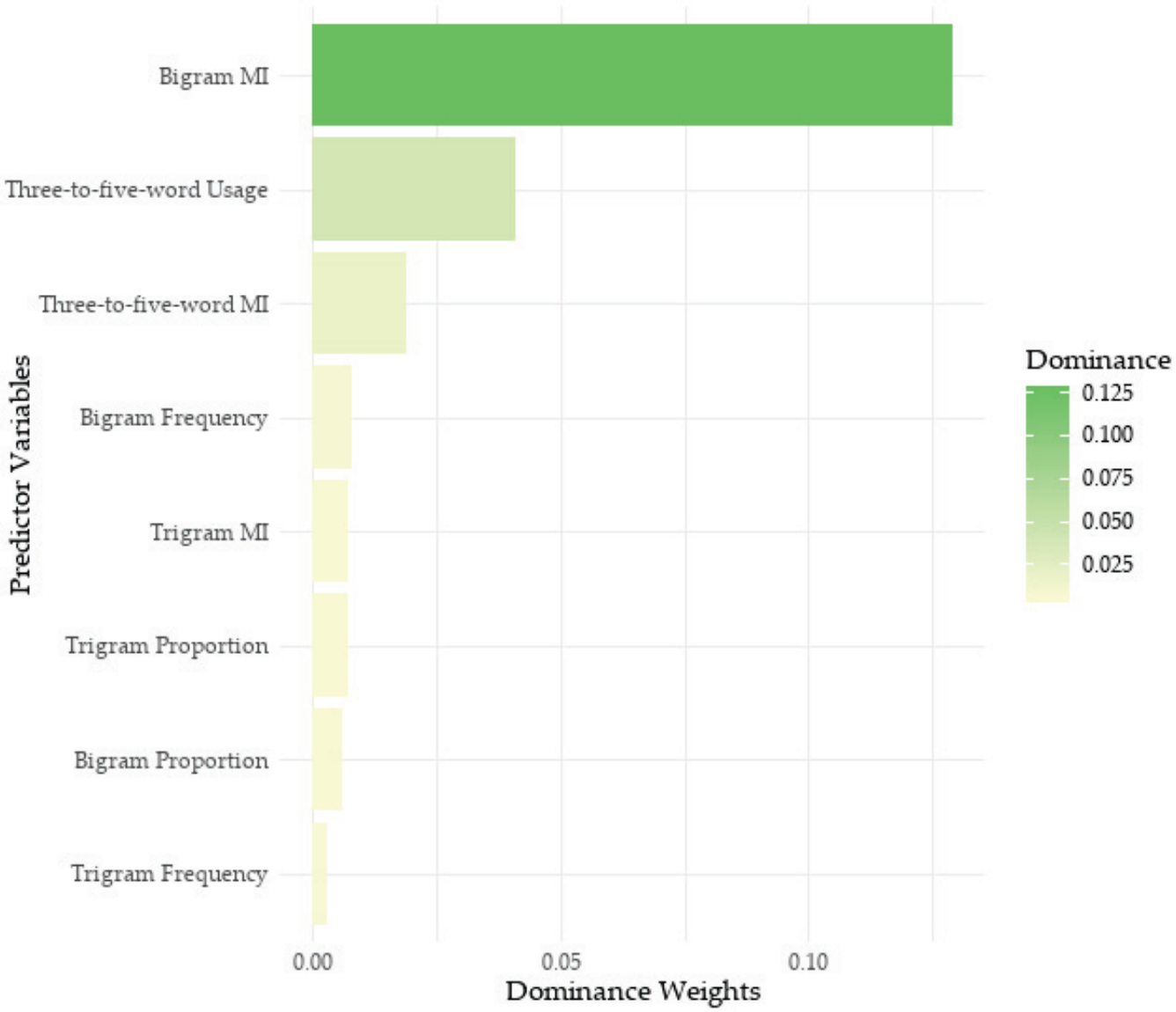

Figure 2. Dominance weights in descending order (criterion: presentation raw scores).

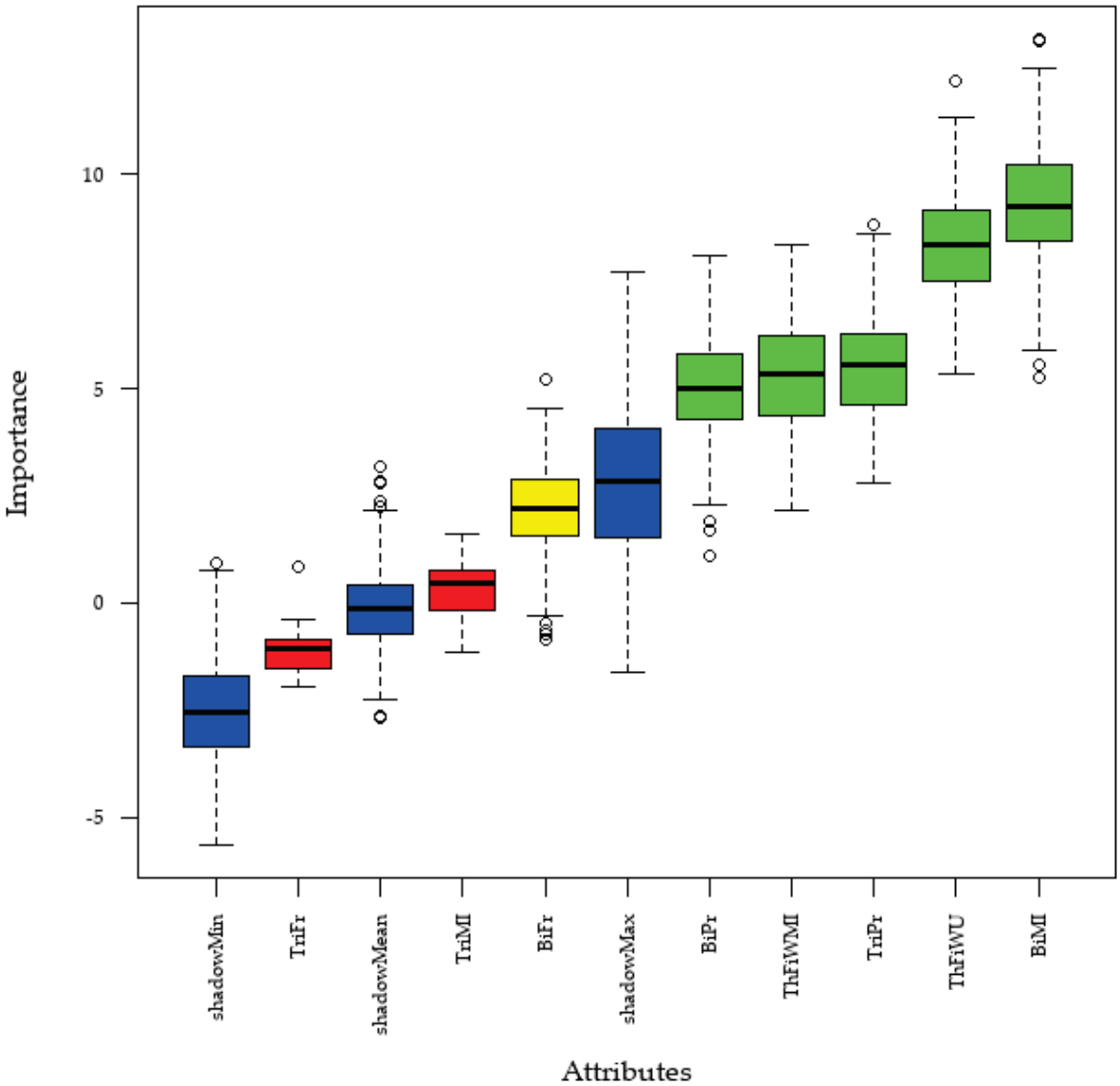

Figure 3. Variable importance plot from random forests using the Boruta algorithm (criterion: presentation raw scores). *Note*: In the Boruta algorithm box plots, green indicates "confirmed" variables, red indicates "rejected" variables, yellow indicates "tentative" variables, and blue indicates "randomized shadow" variables, which serve as a reference to assess the importance of the original variables against random chance.

6. Discussion

The overarching aim of this study was to explore the impact of LB usage on oral proficiency across a broad dataset and to employ a novel LB refinement technique for a more detailed analysis. Our research question asked to what extent LB usage of varying lengths could predict raw presentation scores. The following discussion has been structured to address this research question in relation to theoretical frameworks, previous findings, and our own hypothesis.

The robust regression and dominance analysis identified bigram MI as a significant positive and dominant predictor of raw presentation scores (b = 14.38, 95% CI [8.01, 20.76], t = 4.42; dominance weight = 58.63%). This was complemented by the finding that three-to-five-word LB usage was also found to be a significant positive and powerful predictor of presentation success (b = 0.53, 95% CI [0.06, 0.99], t = 2.22; dominance weight = 18.80%). These two LB measures' significances were reinforced by the variable importance plot

generated by the Boruta algorithm, showing consistency across different approaches. Our discussion will thus focus on these two LB measures.

Our finding that learners who effectively use text-external high MI bigrams achieve higher proficiency levels are consistent with Pawley and Syder's (1983) lexical phrases theory. Pawley and Syder argued in their seminal work that the native-like selection of expressions, which includes collocations and idiomatic phrases captured by measures like bigram MI, is an integral aspect of speaking a language fluently. Their lexical phrases theory posits that a speaker's proficiency is marked by the ability to produce sequences of words that native speakers recognize as familiar, suggesting that native-like proficiency is to some extent a function of memory for lexically stored sequences rather than just rules for combining words. Our findings support and extend this notion by suggesting that not just the presence of LBs, but their "rare exclusivity", which is the main practical effect of the MI score (Gablasova et al. 2017, p. 10), differentiates more proficient speakers from their less proficient counterparts. The empirical evidence presented in this paper supports the view that language proficiency, especially in productive skills such as speaking, involves the use of language patterns that are characteristic of L1 speaker usage. Our findings suggest that as L2 proficiency increases, so does the use of high MI score bigrams, indicating a narrowing gap in collocational usage between L2 and L1 speakers. These finding highlight the importance of integrating specific types of LBs, particularly those with high MI, into language assessment and instruction, supporting the idea that proficiency is not merely a matter of lexical range but also involves the strategic use of linguistically sophisticated and exclusive lexical patterns.

Our results also align with Kormos's (2006) extension of Levelt's model to L2 speakers by illustrating the formulation stage's critical role in speech production. Specifically, the increase in text-external bigram MI and longer text-internal LB usage with proficiency suggests a more efficient lexical selection process among higher proficiency learners, echoing the notion that advanced speakers can activate and employ appropriate lemmas with greater ease. This efficiency likely contributes to the freeing up of cognitive resources for other processing needs, which is essential for achieving fluency. It provides empirical support for the theory that a larger MWS repertoire enables L2 learners to enjoy a processing advantage by reducing the demands on cognitive resources, thereby facilitating more fluent and sophisticated language production. In addition, our findings that bigram MI and longer (three-to-five-word) LB usage predicts higher scores echoes the tenets of usage-based theory, specifically as it relates to language proficiency. Usage-based theory suggests that language learning is exemplified through the increased use and understanding of recurrent patterns or constructions in language, which are acquired through exposure and use (Bybee 2006; Ellis 2002). Bigram MI within the usage-based framework can be seen as a proxy for the type of patterned use that is a hallmark of language proficiency. Bigram MI measures the exclusivity and specificity of the co-occurrence between two words, reflecting how their joint appearance significantly exceeds what would be anticipated based on their independent distribution across texts. This measure not only highlights word pairs that share a unique connection but also underscores the meaningful combinations that are preferentially utilized by more proficient speakers. The high bigram MI scores in our study suggest that more proficient language users are more likely to employ exclusive and information-rich lexical sequences. In a sense, bigrams with high MI scores are those that are entrenched in the linguistic repertoire of proficient speakers, thus reflecting common usage patterns.

Our findings, highlighting the predictive power of bigram MI for proficiency levels, are also consistent with key insights from corpus-based SLA studies that have explored collocational usage differences between L1 and L2 writers. These findings, while encouraging, are for writing, while our paper considers speaking. Notably, the research in this area has consistently found that L1 users tend to produce collocations with higher MI score values compared to L2 users (Durrant and Schmitt 2009; Ellis et al. 2015; Schmitt 2012). The work of Schmitt (2012, p. 6) emphasizes this disparity, concluding that the absence of

high MI collocations is a distinctive marker of non-native versus native production. Our findings also echo the findings of previous research, which compared bigram usage among L2 writers across different proficiency levels. Granger and Bestgen (2014) compared the use of collocations by intermediate and advanced L2 writers, finding that essays scoring in the advanced range of the CEFR had higher proportions of bigrams with high MI scores than essays scoring in the intermediate range. These convergent findings highlight the significance of the MI score as a measure for distinguishing between the collocational choices of L1 and L2 users, as well as between intermediate and advanced L2 users, thus highlighting its utility and growing importance in language research. The differential use of high MI score collocations between L1 and L2 users and between intermediate- and higher-proficiency L2 users highlights a key aspect of language proficiency: the ability to employ exclusive, less frequent word combinations beyond common collocations.

However, a notable divergence arose concerning our bigram MI finding and the other existing literature focusing on spoken production. Notably, Tavakoli and Uchihara (2020) found a general decrease in MI with rising proficiency, suggesting a broadening between LB combinations as learners become more proficient. Conversely, our results showed an increase in bigram MI with proficiency levels (except for a slight decrease in trigram MI). Although both studies used TAALES software to measure bigram MI through a text-external approach, the difference in findings could be due to the distinct data sets, methodologies, or proficiency level groupings employed in the two studies. While Tavakoli and Uchihara used a Kruskal–Wallis H test because of violations of the homogeneity of variance, our study applied robust regression to account for outliers, which may have contributed to the contrasting outcomes in the MI trends. While both studies corroborate the trend of increased LB usage with proficiency, the current analysis contributes a novel perspective by documenting the pattern of bigram MI increase and by bringing to light the intricate usage of longer LBs as learners progress. Such contrasting results underscore the need for further research to explore the intricate dynamics of bigram MI use and proficiency levels, hopefully enriching our understanding of effective language use by L2 speakers.

Regarding our hypothesis that the use of LBs of all lengths would positively correlate with raw oral presentation scores, the results show a mixed but insightful picture. The robust regression and dominance analysis, corroborated by the Boruta algorithm, strongly support the hypothesis in the case of bigram MI and three-to-five-word usage. Bigram MI, which emerged as the most powerful and dominant factor, accounted for a significant portion of the variance in presentation scores. This likely reflects the importance of conciseness in effective oral communication, where the use of meaningful and compact word pairs (bigrams with high mutual information) enables speakers to convey their points succinctly, thereby engaging the audience effectively. The positive influence of three-to-five-word usage also highlights the value of specificity. These slightly longer LBs, which we measured through the text-internal approach, are critical for articulating complex or specific ideas in a clear and focused manner, enhancing the speech's informativeness without becoming overly wordy. However, our hypothesis found less support with other types of LBs. Bigram proportion, for instance, showed a negative nonsignificant association with presentation scores, and trigram-related measures (frequency and MI) had lesser influence. This suggests that, while certain LBs positively correlate with presentation performance, not all LB types show the same level of predictive power. These findings partially validate our hypothesis, thus underscoring the significant predictive power of specific types of LBs, particularly bigram MI, in relation to raw presentation scores.

In summary, while our hypothesis that all lengths of LBs would positively predict oral proficiency scores was only partially supported, the current study contributes insights into the specific lexical features (bigram MI in particular) that are most predictive of higher proficiency scores. The current study's findings enhance the existing literature by providing a nuanced picture of how LB usage evolves at higher levels of language learning. Our findings also contribute to research methodology by showing the effectiveness of combining robust statistical techniques with machine learning algorithms like Boruta to

strengthen the robustness of educational research. The convergence of evidence from different analytic techniques strengthens the reliability of the current study's findings. Using robust regression helped to mitigate the influence of outliers, dominance analysis provided insights into the relative importance of predictors, and the Boruta algorithm offered an additional layer of confirmation about which LB attributes are truly influential. Such methodological triangulation enhances the study's credibility, allowing for more confident conclusions regarding the predictive power of LB usage for speaking proficiency.

7. Limitations and Suggestions for Future Research

Although the results are encouraging, there is still room for future research to overcome the current study's limitations. One notable area that was not explored in the current study is the potential influence of different presentation topics on LB usage and presentation scores. This unexplored area might shed further light on findings such as the increased use of high MI score bigrams among higher-scoring learners. For example, Gablasova et al. (2017) emphasize that the MI score favors less frequent and more specialized combinations, such as technical terms, which can vary significantly with the nature of the presentation topic. They note that "the technical nature of a specific topic can influence the strength of collocations as measured by the MI score", often revealing hidden patterns when only generalized MI score rankings are considered across whole corpora. (Gablasova et al. 2017, p. 20). This observation could be highly pertinent to our research, indicating that the frequency and variety of high MI score word pairs we have observed might not only reflect the speakers' language proficiency but also the specific vocabulary requirements of their chosen topics. Technical presentations, for instance, might necessitate the use of specialized vocabulary, thereby increasing the MI scores of the collocations used. Future studies could benefit from incorporating an analysis of the presentation topics, examining how the choice of topic influences the use of high MI score bigrams across proficiency levels.

At least five other potential limitations warrant consideration. First, the proficiency level of the participants was confined to intermediate- and higher-proficiency learners (IELTS bands 6.5 to 7.5), thus excluding lower-proficiency speakers and potentially limiting generalizability. Second, the study's cross-sectional design limits the ability to trace language proficiency development over time or establish causality between variables. A longitudinal design could be useful in future studies, monitoring LB usage and proficiency over time. Third, the frequency-based approach using corpus analysis software has certain limitations, especially when dealing with spoken corpora. The current study used a minimum frequency of three and a minimum range of three to identify LBs in the learner corpus in order to keep the analysis manageable. These minimum frequency cutoffs mean that this study did not identify all multiword units in a comprehensive way. Some multiword units were used only once or twice or used idiosyncratically or in a nonstandard way by the L2 learners in our corpus. Many multiword sequences tend to blend into the linguistic context in transcripts, and many are frames or have larger fillable slots, which present real challenges for automatic extraction techniques. Such infrequent and/or semifixed units were not detected in the current study's approach. The current study focused on bi- and trigrams using TAALES and three-to-five-word phrases using AntConc, possibly leaving out other multiword structures that may affect speaking proficiency. Future research could expand the scope of MWS analysis for more comprehensive insights. Fourth, the operationalization of LB usage through MI and frequency does not account for the qualitative aspects of contextual appropriateness in conversation. Lastly, since MI scores were primarily designed for two-word collocations, and since they do not consider the order of the words (Biber 2009; Hyland 2012), the MI scores might not reliably measure longer lexical strings. Future research should aim to address these limitations.

8. Pedagogical Implications and Suggestions

The findings support including bigram MI and longer LBs in assessing and teaching English proficiency, especially in speaking. They highlight the importance of focusing on

the quality of LBs, particularly high MI bigrams, rather than just their quantity. This insight can inform instructional design to enhance learners' speaking proficiency. Here are a few practical steps for implementing these findings in the classroom:

1. Select relevant texts: Choose texts that align with students' interests and proficiency levels.
2. Identify high MI bigrams: Use Tom Cobb's "Phrase Extractor" tool (https://www.lextutor.ca/multiwords/phrase/, accessed 3 February 2024) to extract high MI bigrams from the selected texts. These bigrams often serve as the foundation for longer LBs and expressions, thus making them a useful starting point.
3. Practice and raise awareness:
    a. Develop exercises: Create exercises that target high MI bigrams and longer LBs.
    b. Highlight during activities: Have students notice and highlight high MI bigrams and longer LBs during reading and listening activities.
    c. Assessment criteria: Include criteria related to use of high-quality LBs in speaking (and writing) grading rubrics.

By integrating and building on these steps, educators can effectively enhance students' proficiency and awareness of high-quality LBs in English.

9. Conclusions

The current study has provided detailed insights into the patterns of LB usage that correspond to higher oral presentation scores. By extending beyond the scope of Tavakoli and Uchihara (2020), it has illuminated the intricate relationship between proficiency and both the frequency and complexity of LBs. Through extensive statistical analysis, it has established the link between linguistic sophistication and presentation scores, emphasizing the importance of both shorter and longer LBs in speaking proficiency. The current study's findings make a persuasive case for the critical role of LB usage, particularly bigram MI, in predicting English language proficiency and presentation performance. These findings enhance our understanding of the relationship between lexical choice and speaking performance, thus offering practical insights for language assessment and teaching. The current research has highlighted bigram MI (and to some extent three-to-five-word usage) as a reliable indicator of English language scores on oral academic presentations. This is consistent with linguistic theories first put forward in the 1980s (e.g., Pawley and Syder 1983; Levelt 1989) that emphasize the efficient use of preformed chunks of language as a requirement of fluent speech and therefore a hallmark of proficiency. The implications for language teaching are considerable, suggesting that educators should include a focus on teaching strategies that improve learners' awareness and command of high MI bigrams.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/languages9070232/s1; Table S1. Assumption checks for regression models; Table S2. Variance inflation factor (VIF) values; Figure S1. Cook's distance plots for detecting influential observations in the regression model of presentation raw scores; Table S3. R code and detailed results of the dominance analysis for presentation raw scores; Table S4. R code and detailed results of random forests and Boruta analysis for presentation raw scores.

Author Contributions: Conceptualization, D.H., J.C., T.U. and G.H.; Data curation, D.H. and G.H.; Formal analysis, D.H.; Funding acquisition, D.H., J.C., T.U. and G.H.; Investigation, D.H., J.C., T.U. and G.H.; Methodology, D.H., J.C., T.U. and G.H.; Project administration, D.H., J.C. and G.H.; Resources, D.H., J.C., T.U. and G.H.; Supervision, J.C.; Validation, D.H.; Visualization, D.H.; Writing—original draft, D.H.; Writing—review & editing, D.H., J.C., T.U. and G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by two Grants-in-Aid for Scientific Research (No. 21K00669 and No. 22K00700) from the Japan Society for the Promotion of Science. The authors are very grateful for this support.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and APA ethical standards. It was approved by the Ethics Committee of Queen Mary University of London where the data collection took place (research ethics approval number: QMREC2414a).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We extend our sincere gratitude to the two anonymous reviewers and the *Languages* production team for their valuable comments and suggestions. Their insightful feedback has greatly improved the quality of our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Presentation grading descriptors.

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Presentation Content [20]	• Purpose of presentation is clear, appropriate, and fully achieved. • Presentation is clearly focused, and only relevant issues presented. • Excellent research, which is clearly demonstrated through illustrations and examples. • All source material cited. • Visual aids are designed to a professional standard in terms of layout, bibliography, and contents. • Very good analysis, synthesis, and application of research.	• Purpose of presentation is clear, appropriate, and fully achieved. • Presentation is focused, and only relevant issues presented. • Appropriate research is clearly demonstrated through illustrations and examples. • All source material cited. • Clear, well-designed visual aids; effectively proofread. • Good analysis, synthesis, and application of research.	• Purpose of presentation is clear, appropriate, and largely achieved. • Presentation focused; issues presented are mainly relevant issues. • Appropriate research is demonstrated through illustrations and examples. • All source material cited, despite minor errors. • Generally clear and well-designed visual aids. Some evidence of proofreading, but some errors may persist despite this. • Some evidence of ability to analyze, synthesize, and apply research.	• Appropriate and adequately achieved purpose, though may lack clarity. • May be occasional loss of focus and irrelevancies in parts. • Presentation shows some evidence of research and an understanding of the topic. • All source material is cited, though with some errors. • Generally satisfactory design of visual aids. Some lack of proofreading may result in careless mistakes. • Presentation may be more descriptive than analytical.	• Purpose of presentation is appropriate but may not be entirely achieved. • Some loss of focus and some irrelevancies may be evident. • Presentation demonstrates evidence of adequate research and some understanding of the topic • Most source material is cited, though with frequent errors. • Adequately designed visual aids. Inadequate proofreading may lead to careless mistakes. • Presentation may be rather descriptive.	• Purpose of presentation may be unclear or inappropriate. • Presentation is generally unfocused with many irrelevancies. • Presentation demonstrates little evidence of research and weak understanding of the topic. • Some citation of source material. • Visual aids may provide inadequate support for the presentation. Inadequate proofreading or lack of proofreading may result in careless errors. • Presentation may be largely descriptive.	• Purpose of presentation unclear or inappropriate. • Presentation is unfocused and contains many irrelevancies. • Presentation demonstrates no evidence of research and limited understanding of the topic. • Little or no citation of source material. • Visual aids nonexistent or inadequate. Lack of proofreading results in incom-prehension. • Presentation may be entirely descriptive.

Table A1. *Cont.*

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Presentation Structure [20]	• Excellent flow; causes no difficulties for listener. • Logical sequencing of ideas and very good organization of presentation. Excellent introduction and conclusion. Questions invited. • Very good organization within sections. • Excellent use of signposting expressions to create cohesion and coherence.	• Good flow; causes no difficulties for listener. • Logical sequencing of ideas good organization of presentation. Very clear introduction and conclusion. Questions invited. • Good organization within sections. • Effective use of signposting expressions to create cohesion and coherence.	• Reasonably good flow; causes few difficulties for listener. • Good sequencing of ideas, which enables the message to be followed clearly. Good introduction and conclusion. • Reasonably good organization within sections, although some room for improvement. Questions invited. • Fairly good use of signposting language to create cohesion and coherence.	• Reasonable flow causes occasional difficulties for listener. • Some ability to sequence ideas, but overall structure may contain flaws. Reasonable introduction and conclusion. • Fairly good attempt to organize sections into main and supporting ideas; some use of examples but insufficient. Questions invited. • Attempt at signposting language; sometimes inappropriate or inaccurate; parts may lack cohesion.	• Adequate flow, but causes some difficulties for listener. • Limited ability to sequence ideas, and overall organization may be flawed, but the message can be followed adequately. • Introduction and conclusion may be simplistic, overlong, or rushed. Questions not immediately invited. • Sections may lack unity, but may show an attempt to use topic and supporting sentences. • Attempt at signposting language, but it may be inappropriate; there may be some lack of cohesion.	• Lack of flow but causes strain for listener. • Flawed overall sequence of ideas, but message can be followed in places. Introduction and conclusion: simplistic, weak, do not correspond to body. Questions not invited. • Section structure may be weak and disconnected—sections short and disjointed; little use of examples and illustrations. • Limited use of signposting language and often inappropriate; some lack of cohesion.	• Absence of flow, which often puts strain on listener. • Ideas are poorly sequenced and organized, and the message is difficult to follow. Lacks clear organization structure. Little understanding of the purpose of introductions and conclusions. Questions not invited. • Sections poorly organized and show little understanding of the purpose of structure. • Very limited or inaccurate use of signposting language; lack of cohesion.
Seminar Leadership [20]	• There is a totally clear task for seminar participants, and the content is all highly focused and relevant. • The student clearly demonstrates a very high level of awareness of his/her audience. • The discussion is excellently controlled throughout. • The student gives a highly lucid summary of the discussion at its conclusion.	• There is a clear task for seminar participants, and the content is focused and relevant. • The student demonstrates very good awareness of his/her audience. • The discussion is very well controlled. • The student gives a very good, lucid summary of the discussion at its conclusion.	• There is a fairly clear task for seminar participants, and the content is mostly focused and relevant. • The student demonstrates good awareness of his/her audience. • The discussion is well controlled. • The student gives a good summary of the discussion at its conclusion.	• There is a task for seminar participants, and the content is mostly relevant, but there may be some lack of clarity. • The student has satisfactory awareness of his/her audience. • An acceptable attempt is made to control the discussion. • The student gives a satisfactory summary of the discussion at its conclusion.	• There is a task for seminar participants, but it may not be presented clearly. Some of the content may lack focus and relevance. • The student may lack awareness of his/her audience. • The discussion may not be well controlled. • The student gives a summary of the discussion at its conclusion, but this may lack clarity.	• There may be some confusion about the task for seminar participants. The content lacks focus and relevance. • The student lacks awareness of his/her audience. • The discussion is only just controlled. • The student gives a summary of the discussion at its conclusion, but this lacks clarity.	• The task for seminar participants may be inappropriate or unclear and is poorly explained. The content is unfocused and irrelevant. • The student has little or no awareness of his/her audience. • The discussion is not controlled. • The student fails to give a summary of the discussion at its conclusion or does this very poorly.

Table A1. *Cont.*

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Language Fluency [20]	• Clear pronunciation all the time. • Very good, fluent command of language, with almost no hesitations and excellent control of speed. • Excellent use of intonation and stress to convey stance and topic changes • Register always appropriate for this type of interaction. • Script independent; very confident and effective use of nonverbal communication (e.g., facial expressions, appropriate appearance);	• Clear pronunciation most of the time. • Good, fluent command of language, with few hesitations and very good control of speed. • Very good use of intonation and stress to convey stance and topic changes • Register always appropriate for type of interaction. • Script independent; confident and effective use of nonverba communication (e.g., facial expressions, appropriate appearance).	• Generally clear pronunciation. • Good, fluent production, with some hesitations but good control of speed. • Generally good use of intonation and stress to convey stance and topic changes. • Register generally appropriate for type of interaction. • Generally, script independent; effective use of nonverbal communication (e.g., facial expression, appropriate appearance).	• Pronunciation is clear, but there are some mispronunciations. • Speaks with a degree of fluency, but with limited control of speed and some hesitations. • Reasonable use of intonation and stress to convey topic changes, but stance may not always be evident. • Register reasonably appropriate for type of interaction. • Often script independent; often effective use of non-verbal communication (e.g., facial expressions) and acceptably appropriate appearance.	• Pronunciation is generally clear enough to be understood despite a noticeable accent. • Can speak, but with significant hesitation. May require a 'sympathetic' interlocutor. • Intonation and stress may only occasionally be used to convey stance or topic change. • Register is just appropriate; may sometimes be inappropriate for type of interaction. • Partly script independent; some limited awareness of nonverbal communication (e.g., facial expressions used effectively on occasion, fairly appropriate appearance).	• Mispronunciation sometimes makes communication difficult. • Hesitations can make communication difficult. Speed may be too fast or too slow. Often requires a 'sympathetic' interlocutor. • Stance and topic change not signalled with intonation and stress. • Register is often inappropriate for interaction. • Script dependent; little awareness of nonverbal communication (e.g., facial expressions sometimes inappropriate, fairly inappropriate appearance).	• Mispronunciation severely impedes communication. • Frequent hesitation or lack of control over speed severely impedes communication. Requires a 'sympathetic' and active interlocutor.' • Little control of intonation and stress. • Register is inappropriate for interactions. • Script dependent; poor awareness of nonverbal communication (e.g., inappropriate facial expressions and/or inappropriate appearance).

Table A1. *Cont.*

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Language Accuracy [20]	• Student demonstrates mastery of the grammar required for the task; excellent ability to manipulate complex structures. • Excellent use of vocabulary, which is appropriate to the task. • Excellent academic style, with totally appropriate use of register, very good ability to express caution and to avoid overgeneralizing. • Clear evidence of proofreading (in visuals) and practice in presentation.	• Student demonstrates an authoritative use of the grammar required for the task; good ability to manipulate complex structures. • Good use of vocabulary, which is appropriate to the task. • Very good academic style with appropriate use of register, good ability to express caution and to avoid overgeneralizing. • Clear evidence of proofreading (in visuals) and practice in presentation.	• Student shows an above average level of use of grammar required for the task; some use of complex structures but perhaps incorrect use. • Good range of appropriate vocabulary. • Good awareness of academic style (register, expression of caution, few over-generalizations). • Good evidence of proofreading (in visuals) and practice in presentation, but some errors may persist despite this.	• Student shows a reasonable use of grammar with some ability to manipulate complex structures. There may be a limited number of grammatical errors, but these do not interfere with meaning. • Vocabulary generally appropriate to the task. • Awareness of academic style, but some inappropriate register; expression of caution may be weak, and overgeneralizations may be evident. • Some lack of proofreading (in visuals) and practice in presentation may result in careless mistakes.	• Student shows a basic grasp of grammar, but limited ability to manipulate complex structures. Errors may interfere with meaning. • Adequate range of appropriate vocabulary: a narrow range of simple language. • Some awareness of academic style, but there are likely to be several overgen-eralisations and limited ability to express caution. • Inadequate proofreading (in visuals) and practice in presentation may lead to careless mistakes.	• There may be recurrent grammatical errors and limited ability to manipulate com-plex structures. • Some inappropriate use of vocabulary. • Choice of style and register is often inappropriate. • Inadequate proofreading and practice may result in careless errors.	• Significant, recurrent grammatical errors. Very limited ability to manipulate structures appropriately and frequent errors in basic grammati-cal structures. • Range of vocabulary is inadequate for the task; errors make the meaning difficult to discern and cause strain for the reader. • Limited or no ability to use academic style. • Lack of proofreading and practice results in incomprehen-sion.

Note. These descriptors were used by teachers who assessed the presessional presentations at Queen Mary University of London. Shared with permission from Queen Mary University of London.

Appendix B

Table A2. The most frequent three-, four-, and five-word recurrent word combinations used by all speakers (N = 150): Ranked according to MI score.

Rank (FREQ)	Rank (MI)	Bundles	Freq	Range	FREQ as 5-Gram	MI Score
21	1	if you have any questions	22/18 *	22/18	18	33.14
29	2	market forces and free competition (are)	18/18	9/9	18	32.59
73	3	in a socially responsible way	6/5	5/4	5	32.17
98	4	the possibility of armed conflict	4/4	3/3	4	28.58
111	5	Let's start by taking	3	3		25.07
63	6	I shall(will) only take	7(3)	7(3)		24.89
56	7	the presentation will be brief	10/10	10/10	10	23.62
34	8	(seven) minutes of your time	14	14		23.40
101	9	a wide range of	3	3		22.34
78	10	have a seminar discussion	5	5		21.27
77	11	does globalization lead to	5	5		20.87
37	12	at the same time	12	12		20.78
47	13	as we all know	11	11		20.65
95	14	can be defined as	4	3		20.50
69	15	will last about seven (to eight minutes)	6	6		20.43
86	16	(have/has) a positive impact on	5	4		20.38
97	17	for a long time	4	3		20.32
108	18	it can be seen	3	3		20.10
99	19	a large amount of	3	3		20.06
76	20	the rich and the poor	4/4	4/4	4	19.99
24	21	as you(we) can see (from)	20(7)	17(7)		19.87
80	22	(to) help you understand the	5	5		19.72
102	23	advantage and disadvantage of	3	3		19.71
74	24	on the other hand	6	5		19.55
2	25	I would like to (talk about the/discuss/explain/introduce)	50	39		19.50
61	26	all over the world	8	6		19.31
75	27	we all know that	6	4		19.26
96	28	developed and developing countries	4	3		18.79
100	29	a large number of	3	3		18.72
5	30	(firstly/then) I will talk about	37	24		18.49
118	31	to deal with this	3	3		18.26
85	32	(my) presentation is(will be) divided into	5(4)	5(4)		18.25
50	33	I will focus on	10	10		18.24
117	34	This talk will last	3	3		17.58
13	35	corporate social responsibility (is)	26	12		17.51
105	36	in the United States	3	3		17.08
43	37	we are going to (talk)	12	6		16.76
109	38	it is easy to	3	3		16.32
110	39	my presentation today is	3	3		16.21
9	40	(is) my thesis statement (is/and)	28	22		16.04
67	41	Let's turn to the	7	5		15.88
25	42	at the end of (my/the talk/presentation)	20	19		15.86
88	43	as a result of	4	4		15.58
92	44	I will make a	4	4		15.27
40	45	(I will be) glad(happy) to answer (them)	12(5)	12(5)		15.18
106	46	is a form of	3	3		14.89
58	47	the last one is (the)	8	8		14.79
23	48	and then I(we) will (show)	21(4)	19(4)		14.70
116	49	there will be a	3	3		14.60
107	50	is the most important	3	3		14.39
119	51	will move on to	3	3		14.20
82	52	(I) will start with the	5	5		14.16
87	53	moving on to the	5	4		14.15
10	54	Let's move on (to the)	28	22		14.10
81	55	I will analyze the	5	5		14.00
89	56	in the context of	4	4		13.95
112	57	Let's talk about the	3	3		13.90
1	58	(today/firstly/then) I'm going to (talk about/present/analyze)	52	34		13.87
45	59	Let's start (begin) with (the)	11(5)	11(5)		13.70
114	60	the last part is	3	3		13.24
113	61	So in this presentation	3	3		13.09
30	62	we can see (the/that)	18	11		12.87
14	63	(I have divided) my presentation(talk) into (four/five parts)	25(4)	25(4)		12.78

Table A2. *Cont.*

Rank (FREQ)	Rank (MI)	Bundles	Freq	Range	FREQ as 5-Gram	MI Score
94	64	So what are the	4	4		12.64
103	65	in the era of	3	3		12.53
104	66	in the field of	3	3		12.40
66	67	a brief introduction (of)	7	7		12.39
18	68	(is/about) the relationship between (a and b)	24	14		12.30
8	69	(first/then/second) I will introduce (the/my/some)	29	18		12.13
4	70	(firstly/and finally) I will give (you/the/a/some)	41	31		11.83
91	71	we move to the	4	4		11.81
93	72	(be) seen as a	4	4		11.66
90	73	that globalization is a	4	4		11.32
65	74	as well as (the)	7	7		11.15
6	75	(Let's) look at the (first part)	34	28		10.97
12	76	I want to (talk about/show)	26	20		10.88
39	77	(this/here) is my outline (and)	12	12		10.82
33	78	(with/about) the background information	15	10		10.78
52	79	(then) I will discuss (the)	10	10		10.75
31	80	We need to (look at)	17	9		10.74
41	81	Let's move to (the)	12	10		10.74
84	82	(the) third one is	5	5		10.51
20	83	first of all (I/the)	23	19		10.38
62	84	in this slide (I)	8	6		10.33
46	85	(and) after that I (will)	11	11		10.24
38	86	(becoming) more and more	12	12		10.12
16	87	(in/and) the second part (is/I)	24	22		10.00
26	88	So what is (globalization/the)	20	19		9.91
42	89	to give you (some/a)	12	10		9.80
51	90	(and) the last one (is the)	10	10		9.69
72	91	to start with (I)	6	6		9.62
115	92	the needs of the	3	3		9.48
3	93	(on to/at) the first part(one) (is/the introduction)	41(12) **	31(10)		9.39
59	94	(and) the third part (is)	8	8		9.33
48	95	we will have (a)	11	10		9.30
7	96	(the outline/focus/part) of my presentation (is)	31(4)	29(4)		9.27
79	97	(be) responsible for the	5	5		9.06
70	98	(will) show you the	6	6		8.79
32	99	in this presentation (I will)	15	13		8.43
54	100	in my presentation (I)	10	10		8.39
64	101	(and) the second one (is)	7	7		8.35
35	102	this presentation is (the)	14	13		8.08
60	103	in this part (I will)	8	7		7.99
49	104	(that) I think the	11	9		7.78
53	105	The purpose of (this presentation is)	10	10		7.68
71	106	the effect of (globalization)	6	6		7.38
28	107	(the) impact of globalization (on)	22	12		7.30
17	108	(go/move) to the next (slide/part/point)	24	20		7.04
27	109	(in) the process of (globalization)	19	14		6.97
15	110	the definition of (the/globalization/CSR)	25	19		6.81
44	111	the protection of (women's labor rights)	12	8		6.58
83	112	(presentation) is going to	5	5		6.45
68	113	(and) the third is	7	6		6.39
19	114	(is) one of the (main/most)	24	20		5.98
11	115	(lead to/with) the (rapid) development of	27	19		5.81
55	116	(about) the introduction of (globalization)	10	7		5.28
36	117	the first is (introduction)	13	8		4.98
57	118	of the world (economy)	9	8		4.92
22	119	This is the (outline)	22	19		4.72

Note: Contracted forms (e.g., *I'm*) were counted as a single word; for example, *I'm going to* was considered to be a trigram rather than a quadgram. * Multiple-frequency figures listed in column 3 represent the individual frequencies of the four-word sequences that make up the longer five-word structure. ** Frequency of the word combination containing the word in parentheses in the fillable slot.

Appendix C

Table A3. The top-20 MI bigrams produced by the highest-scoring 50 participants.

Rank (MI)	Bigram	MI Score	Freq	Range
1	carbon dioxide	11.36	2	1
2	Kyoto protocol	11.19	2	2
3	Hong Kong	11.16	1	1
4	Saudi Arabia	9.77	1	1
5	infectious diseases	9.75	1	1
6	global warming	9.57	2	1
7	21st century	9.20	1	1
8	sexual harassment	9.06	7	1
9	19th century	8.84	1	1
10	breast cancer	8.65	8	8
11	human beings	8.59	1	1
12	greenhouse gas	8.16	3	2
13	intellectual property	8.16	16	5
14	environmental protection	7.84	1	1
15	wide range	7.81	3	3
16	pharmaceutical companies	7.67	1	1
17	virtual reality	7.58	2	1
18	substance abuse	7.57	1	1
19	gas emissions	7.48	1	1
20	17th century	7.41	1	1

Notes

1 Unlike Tavakoli and Uchihara (2020), we did not use t-scores, because it has recently been indicated that they do not measure association very reliably (Gries 2022).

References

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent Word combinations. In *Phraseology: Theory, Analysis, and Applications*. Edited by Cowie Anthony Paul. Oxford: Oxford University Press, pp. 101–22.

Anthony, Laurence. 2022. AntConc (Version 4.0.5) [Computer Software]. Waseda University. Available online: https://www.laurenceanthony.net/software/antconc/ (accessed on 3 February 2024).

Appel, Randy, and David Wood. 2016. Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly* 13: 55–71. [CrossRef]

Barlow, Michael. 2015. Collocate (Version 2.0) [Computer Software]. Athelstan. Available online: https://athel.com/ (accessed on 3 February 2024).

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14: 275–311. [CrossRef]

Biber, Douglas, and Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the *Toefl iBT*® test: A lexico-grammatical analysis. *ETS Research Report Series* 2013: i–128. Available online: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2013.tb02311.x (accessed on 3 February 2024).

Biber, Douglas, and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–86. [CrossRef]

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Boers, Frank, June Eyckmans, Jenny Kappel, Helene Stengers, and Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10: 245–61. [CrossRef]

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 711–33. Available online: http://www.jstor.org/stable/4490266 (accessed on 3 February 2024). [CrossRef]

Cowie, Anthony Paul. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2: 223–35. [CrossRef]

Dang, Thi Ngoc Yen, Averil Coxhead, and Stuart Webb. 2017. The academic spoken word list. *Language Learning* 67: 959–97. [CrossRef]

Davies, Mark. 2009. The 385+ million word *Corpus of Contemporary American* English (1990–2008+). Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14: 159–90. [CrossRef]

De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3: 59–80. [CrossRef]

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, New Series* 2: 225–46. Available online: https://dial.uclouvain.be/downloader/downloader.php?pid=boreal:75157&datastream=PDF_01 (accessed on 3 February 2024).

Durrant, Phil, and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47: 157–77. [CrossRef]

Ebeling, Signe, and Hilde Hasselgård. 2015. Learner corpora and phraseology. In *The Cambridge Handbook of Learner Corpus Research*. Edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. Cambridge: Cambridge University Press, pp. 207–29.

Ellis, Nick 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: 143–88. Available online: https://psycnet.apa.org/doi/10.1017/S0272263102002024 (accessed on 3 February 2024). [CrossRef]

Ellis, Nick C., Eric Frey, and Isaac Jalkanen. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In *Exploring the Lexis-Grammar Interface*. Edited by Ute Römer-Barron and Rainer Schulze. Amsterdam: John Benjamins, pp. 89–114.

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42: 375–96. [CrossRef]

Ellis, Nick C., Rita Simpson-Vlach, Ute Römer, Matthew B. O'Donnell, and Stefanie Wulff. 2015. Learner corpora and formulaic language in second language acquisition research. In *The Cambridge Handbook of Learner Corpus Research*, 1st ed. Edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. Cambridge: Cambridge University Press, pp. 357–78. [CrossRef]

Gablasova, Dana, Vaclav Brezina, and Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67: 155–79. [CrossRef]

Garner, Jamie, and Scott Crossley. 2018. A latent curve model approach to studying L2 N-Gram development. *The Modern Language Journal* 102: 494–511. [CrossRef]

Granger, Sylviane, and Magali Paquot. 2008. Disentangling the phraseological web. In *Phraseology: An Interdisciplinary Perspective*. Edited by Sylviane Granger and Fanny Meunier. Amsterdam: John Benjamins, pp. 27–49.

Granger, Sylviane, and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching* 52: 229–52. [CrossRef]

Gries, Stefan Th. 2022. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5: 1–33. [CrossRef]

Groom, Nicholas. 2009. Effects of second language immersion on second language collocational development. In *Researching Collocations in Another Language*. Edited by Andy Barfield and Henrik Gyllstad. London: Palgrave Macmillan, pp. 21–33.

Grömping, Ulrike. 2006. Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software* 17: 1–27. [CrossRef]

Hasselgård, Hilde. 2019. Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In *Corpus Linguistics, Context and Culture*. Edited by Michaela Mahlberg and Viola Wiegand. Berlin: De Gruyter, pp. 339–62.

Hougham, Dan, Jon Clenton, and Takumi Uchihara. 2024. Disentangling the contributions of shorter vs. longer lexical bundles to L2 oral fluency. *System* 121: 103243. [CrossRef]

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21. [CrossRef]

Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32: 150–69. [CrossRef]

Kormos, Judit. 2006. *Speech Production and Second Language Acquisition*. London: Routledge.

Kursa, Miron B., and Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36: 1–13. [CrossRef]

Kyle, Kristopher, and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49: 757–86. [CrossRef]

Kyle, Kristopher, Scott A. Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50: 1030–46. [CrossRef] [PubMed]

Larson-Hall, Jenifer. 2015. *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. London: Routledge.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Levelt, Willem J. M. 1992. Accessing words in speech production: Stages, processes and representations. *Cognition* 42: 1–22. [CrossRef] [PubMed]

Liakhovitski, Dmitri, Yegor Bryukhov, and Michael Conklin. 2010. Relative importance of predictors: Comparison of random forests with Johnson's relative weights. *Model Assisted Statistics and Applications* 5: 235–49. [CrossRef]

Liu, Dilin. 2012. The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes* 31: 25–35. [CrossRef]

McGuire, Michael, and Jenifer Larson-Hall. 2017. Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal* 34: 1–25. Available online: https://teslcanadajournal.ca/index.php/tesl/article/download/1271/1106 (accessed on 3 February 2024). [CrossRef]

McGuire, Michael, and Jenifer Larson-Hall. 2021. The contribution of high-frequency multi-word sequences to speech rate and listening perception among EFL learners. *Vocabulary Learning and Instruction* 10: 18–29. Available online: https://vli-journal.org/wp/wp-content/uploads/2022/02/VLI_10_2.pdf#page=22 (accessed on 3 February 2024). [CrossRef]

Mizumoto, Atsushi. 2022. Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning* 73: 161–96. Available online: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12518 (accessed on 3 February 2024). [CrossRef]

Myles, Florence, and Caroline Cordier. 2017. Formulaic sequence (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition* 39: 3–28. [CrossRef]

Nation, Paul. 2013. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2004. What are collocations? In *Phraseological Units: Basic Concepts and Their Application*. Edited by David John Allerton, Nadja Nesselhauf and Paul Skandera. Karlsruhe: Schwabe, pp. 1–21.

Paquot, Magali, and Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32: 130–49. [CrossRef]

Pawley, Andrew, and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and Communication*. Edited by Jack C. Richards and Richard W. Schmidt. London: Longman, pp. 29–59.

Plonsky, Luke, and Deirdre J. Derrick. 2016. A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal* 100: 538–53. [CrossRef]

R Development Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: http://www.r-project.org/ (accessed on 3 February 2024).

Schmidt, Richard. 1992. Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition* 14: 357–85. [CrossRef]

Schmitt, Norbert. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. London: Palgrave Macmillan.

Schmitt, Norbert. 2012. Formulaic language and collocation. In *The Encyclopedia of Applied Linguistics*, 1st ed. Edited by Carol A. Chapelle. Hoboken: Wiley.

Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An academic formulas list (AFL). *Applied Linguistics* 31: 487–512. [CrossRef]

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna, and Stefania Spina. 2020. Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning* 70: 420–63. [CrossRef]

Skehan, Peter. 2014. *Processing Perspectives on Task Performance*. Amsterdam: John Benjamins.

Staples, Shelley, Jesse Egbert, Douglas Biber, and Alyson McClair. 2013. Formulaic sequences and EAP development: Lexical bundles in the TOEFL iBT writing section. *English for Specific Purposes* 12: 214–25. [CrossRef]

Stengers, Helene, Frank Boers, Alex Housen, and June Eyckmans. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics* 49: 321–43. [CrossRef]

Suzuki, Yuichi, Masaki Eguchi, and Nivja de Jong. 2022. Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective. *TESOL Quarterly* 56: 1290–319. [CrossRef]

Tavakoli, Parvaneh. 2011. Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal* 65: 71–79. [CrossRef]

Tavakoli, Parvaneh, and Takumi Uchihara. 2020. To what extent are multiword sequences associated with oral fluency? *Language Learning* 70: 506–47. [CrossRef]

Tremblay, Antoine, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61: 569–613. [CrossRef]

Uchihara, Takumi, and Jon Clenton. 2020. Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research* 24: 540–56. [CrossRef]

Uchihara, Takumi, and Tetsuo Harada. 2018. Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly* 52: 564–87. Available online: https://www.jstor.org/stable/44987081 (accessed on 3 February 2024). [CrossRef]

Uchihara, Takumi, Masaki Eguchi, Jon Clenton, Kristopher Kyle, and Kazuya Saito. 2021. To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech* 65: 311–36. [CrossRef] [PubMed]

Wood, David. 2009. Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics* 12: 39–57. Available online: https://journals.lib.unb.ca/index.php/CJAL/article/download/19898/21737/ (accessed on 3 February 2024).

Wood, David. 2010. *Formulaic Language and Second Language Speech Fluency: Background, Evidence, and Classroom Applications*. Dallas: Continuum.

Wood, David, and Randy Appel. 2014. Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes* 15: 1–13. [CrossRef]

Zhang, Xiaopeng, Baoshan Zhao, and Wenwen Li. 2021. N-gram use in EFL learners' retelling and monologic tasks. *International Review of Applied Linguistics in Language Teaching* 61: 939–65. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Mastery of Listening and Reading Vocabulary Levels in Relation to CEFR: Insights into Student Admissions and English as a Medium of Instruction

Zhiqing Li [1], Janis Zhiyou Li [1], Xiaofang Zhang [1] and Barry Lee Reynolds [1,2,*]

[1] Faculty of Education, University of Macau, Macau SAR, China; hc01131@connect.um.edu.mo (Z.L.); hc01130@connect.um.edu.mo (J.Z.L.); yc17111@um.edu.mo (X.Z.)
[2] Centre for Cognitive and Brain Sciences, University of Macau, Macau SAR, China
* Correspondence: barryreynolds@um.edu.mo

Abstract: Prior to enrolling in an English as a medium of instruction (EMI) institution, students must show an English proficiency level through meeting a benchmark on a standard English proficiency test, which is typically aligned with the Common European Framework of Reference for Languages (CEFR). Along with overall English proficiency, aural/written vocabulary level mastery could also predict students' success at EMI institutions, as students need adequate English vocabulary knowledge to comprehend lectures and course readings. However, aural/written vocabulary level mastery has yet to be clearly benchmarked to CEFR levels. Therefore, this study aimed to investigate the correlations between students' aural/written vocabulary level mastery and their CEFR levels. Forty undergraduate students in a Macau EMI university were recruited to take one English proficiency test and two vocabulary level tests (i.e., Listening Vocabulary Levels Test (LVLT) and the Updated Vocabulary Levels Test (UVLT)). Correlation analyses were conducted to explore the relationship between students' CEFR levels and their mastery of listening and reading vocabulary levels. A positive correlation was found between students' CEFR levels and their mastery of receptive aural vocabulary levels ($\rho = 0.409$, $p = 0.009$). Furthermore, a statistically significant positive correlation was found between students' CEFR levels and their mastery of receptive written vocabulary levels ($\rho = 0.559$, $p < 0.001$). Although positive correlations were observed, no clear pattern was identified regarding the relationship between students' CEFR levels and their mastery of aural/written vocabulary levels. Regression analyses were further conducted to determine the extent to which the combination of receptive aural and written vocabulary knowledge predicts the CEFR levels. The results indicated that the regression model that included only UVLT scores better predicted the CEFR levels. Given the positive correlations observed between students' CEFR levels and their mastery of vocabulary levels, this study's findings suggest the inclusion of aural/written vocabulary levels as additional indicators for ensuring student academic success in EMI institutions. Implications for EMI universities on student admissions, classroom teaching, and provision of additional English courses were provided.

Keywords: vocabulary levels; English as a medium of instruction (EMI); Common European Framework of Reference for Languages (CEFR); student admissions

1. Introduction

The internationalization of education has led to an increase in the implementation of English as a medium of instruction (EMI) (Reynolds and Yu 2018; Schoepp 2018). As a multilingual and multicultural region, Macau has embraced EMI within its higher education system (Reynolds et al. 2023). Specifically, Macau's higher education institutions offer programs in English to enhance their competitiveness as a global education hub (Botha 2013). Moreover, to ensure that students possess adequate language competencies to meet the

academic demands, EMI universities have established English proficiency requirements aligned with the Common European Framework of Reference for Languages (CEFR) as prerequisites for university admissions (Wu 2023). Second language (L2) English-speaking applicants generally need to attain a specific score on standardized international language tests (e.g., IELTS and TOEFL). These tests serve to demonstrate students' proficiency in comprehending and engaging with academic lectures effectively. For example, the University of Macau employs IELTS band 6.0 and TOEFL 79 as the benchmarks for admission to EMI undergraduate programs (University of Macau 2024), which is equivalent to the B2 level on the CEFR scale.

However, a discrepancy has been observed between students' CEFR-defined scores and their actual ability to comprehend EMI lectures and instructional materials (Reynolds et al. 2023). Students have reported linguistic barriers in adapting to the EMI environment (Aizawa et al. 2023; Reynolds et al. 2022), with particular emphasis on the need for reading improvement (Aizawa et al. 2023; Taltz 2011) and extensive vocabulary instruction (Yu et al. 2020). Merely meeting the English proficiency requirements for admissions does not guarantee students' ability to comprehend and actively engage in academic lectures (Reynolds et al. 2023). A previous study conducted in Macau has shown that achieving a 98% lexical coverage in English language teaching lectures requires the acquisition of proper nouns, marginal words, and knowledge of the most frequent 4000 word families (Reynolds et al. 2022). This finding underscored the importance of considering students' vocabulary level mastery when admitting them to various academic programs (Reynolds et al. 2023).

With the increasing implementation of EMI in higher education, it is essential for researchers and educators to consider whether the existing language proficiency tests for university admission are sufficient to examine students' ability to comprehend and engage in EMI lectures. While lexical knowledge is considered fundamental in students' EMI learning outcomes (Uchihara and Harada 2018), a link between students' vocabulary knowledge and their CEFR-defined scores needs to be established to improve the admission process and to ensure that students can excel in an EMI context. Therefore, this study was designed to investigate the relationship between EMI students' CEFR levels and their receptive vocabulary level mastery.

2. Literature Review

2.1. Vocabulary Mastery in EMI Programs

Compared to general English proficiency, vocabulary knowledge is often overlooked when assessing students' readiness for learning in an EMI setting (Reynolds et al. 2023). However, vocabulary knowledge plays a crucial role in determining students' abilities in the four language skills: reading (Uchihara and Harada 2018), writing (Yang et al. 2019), listening (Reynolds et al. 2022, 2023), and speaking (Enayat and Derakhshan 2021). Specifically, vocabulary knowledge is identified as the main challenge students face in an EMI setting (Macaro 2022), as they struggle to understand content materials filled with unfamiliar words and face difficulties in incidental vocabulary acquisition (Hu et al. 2021; Nguyen 2021). These challenges encountered by students in comprehending EMI lectures highlighted the necessity of assessing their receptive aural/written vocabulary knowledge.

Numerous studies have addressed a positive relationship between L2 learners' lexical coverage and their L2 reading comprehension. For instance, the mastery of the most frequent 1000 word families can reach 88% lexical coverage of English proficiency test passages (Stoeckel et al. 2021; Webb et al. 2017). The demand for lexical coverage in comprehending academic lectures is even higher. For example, Dang and Webb (2014) found that achieving 96% coverage of EMI academic lectures requires knowledge of the most frequent 4000 word families, including proper nouns and marginal words. Furthermore, achieving a 98% coverage necessitates knowledge of the most frequent 8000 word families, along with proper nouns and marginal words. These studies highlighted the need to measure L2 learners' vocabulary mastery levels when examining their readiness for learn-

ing in an EMI context. Since vocabulary level mastery offers information on the frequency levels of words that have been mastered by students, knowing their vocabulary level mastery allows for the comparison between students' aural/written vocabulary level mastery and the frequency levels of words required to comprehend EMI lectures and instructional materials from different disciplines. With this information at hand, key stakeholders can more confidently predict students' readiness for learning in specific EMI programs. The Updated Vocabulary Levels Test (UVLT) (Webb et al. 2017) provides insights into the levels of mastery of the most frequent 5000 word families (Webb et al. 2017) and acts as a standard for selecting appropriate vocabulary when designing instructional materials (e.g., Reynolds et al. 2022). Therefore, this study employed the UVLT as a measure of L1 Chinese EFL learners' written receptive vocabulary knowledge, suggesting that a minimum lexical coverage of 90% is required for satisfactory listening comprehension and a higher percentage is beneficial for more comprehensive understanding. However, the explanatory power of vocabulary knowledge in the success of L2 listening comprehension varies from 13% (Wang and Treffers-Daller 2017) to 59% (Cai 2020) among previous studies examining the correlations between L2 vocabulary and listening (Wang 2023). This can be explained by the different measurements of L2 learners' vocabulary knowledge, for example, the format (i.e., a written test versus an aural test) and the focus (i.e., receptive versus productive knowledge) of the vocabulary test. Previous studies focused on orthographic vocabulary size and found a moderate to strong correlation between orthographic vocabulary size and L2 listening comprehension (Bonk 2000; Stæhr 2008; Wang and Treffers-Daller 2017). However, written vocabulary size tests do not directly tap into L2 learners' ability to recognize words in real-time audio input (Matthews and Cheng 2015). Stæhr (2009) emphasized that a vocabulary test involving hearing the target words should be ideally employed when examining the relationship between vocabulary knowledge and L2 listening comprehension. Cheng and Matthews's (2018) study explored the relationship between three measures (i.e., receptive/orthographic, productive/orthographic, and productive/phonological) of vocabulary knowledge and L2 listening comprehension of Chinese tertiary EFL learners. They reported a strong correlation between aural productive vocabulary knowledge and L2 listening comprehension ($r = 0.71$, $p < 0.001$) (Cheng and Matthews 2018). Moreover, Cai (2020) investigated the relationship between L2 learners' receptive aural vocabulary knowledge and L2 listening comprehension, finding L2 learners' receptive aural vocabulary knowledge contributed to the prediction of 59% of the variance in their L2 listening comprehension. Furthermore, McLean et al. (2015) and Wallace (2020) used the LVLT (McLean et al. 2015) to measure Japanese EFL learners' aural receptive vocabulary knowledge and found a significant correlation between their aural vocabulary knowledge and L2 listening comprehension. The validity and generalizability of the LVLT make it an ideal instrument for assessing aural vocabulary knowledge among L2 EFL learners. Du et al. (2022) further confirmed the reliability and validity of the LVLT among Chinese tertiary EFL learners. Therefore, this study used the LVLT as a measure of L1 Chinese EFL learners' aural receptive vocabulary knowledge.

2.2. Mapping Vocabulary Levels to the CEFR

Developed by the Council of Europe, the CEFR provides a description of a foreign language learner's proficiency, utilizing seven descriptors ranging from pre-A1 to C2, in ascending order (Council of Europe 2018). The CEFR is widely recognized as a valuable language evaluation tool used for assessing language proficiency across various languages and contexts (Graves 2008; Hulstijn et al. 2010). The standardized approach of the CEFR enables its users to apply the measurement of individuals' English proficiency, including the four language skills and other aspects, such as vocabulary, grammar, and communicative competence, in a consistent manner (North and Piccardo 2019). Although the CEFR provides general guidelines and a set of descriptors with its focus on reading, writing, listening, and speaking, it cannot refer to any specific language or context of use (Piccardo 2020). Therefore, this makes the utility of supplement testing with appropriate vocabulary

tests (i.e., UVLT, LVLT) particularly compelling. Specifically, this variability underscores the importance of our study: supplementing traditional proficiency tests with specific vocabulary assessments can provide a more comprehensive evaluation of a student's readiness for academic success in an EMI context.

According to the British Council (n.d.), EMI universities have language proficiency requirements that range from B2 to C1, which may vary depending on the educational programs, levels of study, and regions. Due to the varying formats and scales of language tests, it can be challenging to accurately map students' assessment scores to their actual language proficiency (ETS 2010). Therefore, the importance of aligning language proficiency assessments with the CEFR is emphasized. Although the CEFR provides information about the vocabulary range and vocabulary control that L2 learners are expected to have attained at each level of the CEFR hierarchy, the criteria are too general to provide detailed information about the specific vocabulary size associated with each CEFR level (Milton 2010).

Despite the predictive power of students' vocabulary knowledge on their language proficiency, the relationship between CEFR levels and vocabulary requirements has remained ambiguous (Benigno and de Jong 2019). The connections between students' English vocabulary size and the estimated CEFR levels (e.g., Meara and Milton 2003; Milton 2009; Nation, p.c.) are summarized in Table 1. These findings generally showed that a larger vocabulary size is necessary for higher English proficiency, with a vocabulary size of 3000 to 4000 words required for intermediate levels (B1, B2). However, the vocabulary requirements for elementary levels (A1, A2) and advanced levels (C1, C2) vary considerably.

Table 1. Connections between students' English vocabulary size and the CEFR levels.

	X-Lex Score (Meara and Milton 2003)	**X-Lex Score (Adapted from Milton 2009)**	**Suggested Vocabulary Size (Nation, p.c.)** [1]
A1: Beginner	2000 lemmas	<1500 lemmas	120 word families and phrases
A2: Elementary	2750–3240 lemmas	1500–2500 lemmas	The most frequent 1000 word families
B1: Intermediate	3250–3740 lemmas	2500–3250 lemmas	2000–3000 word families
B2: Intermediate	3750–4240 lemmas	3250–3750 lemmas	4000 word families
C1: Advanced	4500–4740 lemmas	3750–4500 lemmas	5000–6000 word families
C2: Native speaker-like	>4750 lemmas	4500–5000 lemmas	7000–9000 word families

Note: [1] The suggested vocabulary size is based on personal communication between the fourth author and Nation through email on 8 April 2024.

Although there have been previous attempts to correlate vocabulary knowledge to the CEFR levels, many of these studies (e.g., Meara and Milton 2003; Milton 2009) have primarily focused on assessing learners' receptive written vocabulary size, overlooking the nuances of their comprehension abilities in different contexts (Milton and Alexiou 2020). Additionally, while a vocabulary size metric is powerful and strongly connected to learners' performance in the four skills (Milton and Alexiou 2020; Stæhr 2008), Siregar and Henni (2023) argued that a vocabulary size test merely provides a general assessment of learners' overall vocabulary knowledge, whereas a vocabulary level test offers more detailed insights into a learner's mastery of vocabulary knowledge. Thus, establishing a link between the CEFR levels and vocabulary level mastery can provide educators and institutions with a clearer understanding of the vocabulary thresholds necessary for learners to progress from one CEFR level to another. Moreover, this linkage caters to the specific language demands of diverse academic and professional contexts, which can be used to inform the language proficiency assessments employed for EMI university admissions. Therefore, this study utilized two vocabulary level tests (i.e., LVLT and UVLT) to examine the relationship between students' aural/written vocabulary level mastery and their CEFR levels.

2.3. Problem Statement and Research Questions

The foci of many previous studies on the relationship between students' CEFR levels and their vocabulary knowledge were limited to vocabulary size, which failed to comprehensively compare students' vocabulary knowledge to the lexical difficulty levels of academic programs. Therefore, a link needs to be established between the CEFR levels and vocabulary level mastery. In addition, most of the previous studies have primarily focused on measuring learners' receptive written vocabulary size (e.g., Meara and Milton 2003; Milton 2009; Milton and Alexiou 2020). Considering that listening comprehension also plays a vital role in understanding lectures for students enrolled in EMI programs, and recognizing a word in its written form does not guarantee recognition of the same word when presented in spoken form (Field 2008; Goh 2000), an equal attention should be given to the measure of receptive aural vocabulary level mastery. Therefore, the current study focused on the relationship between both written and aural receptive vocabulary level mastery and CEFR levels.

Addressing these gaps, the main purpose of this study was to explore the correlation between students' CEFR levels and their mastery of listening/reading vocabulary levels. The following four research questions guided this study:

(1) RQ1: To what extent is the English proficiency of EMI students (i.e., CEFR levels) associated with their mastery of receptive aural vocabulary levels?
(2) RQ2: What correlations exist between EMI students' proficiency (i.e., CEFR levels) and their mastery of receptive aural vocabulary levels?
(3) RQ3: To what extent is the English proficiency of EMI students (i.e., CEFR levels) associated with their mastery of receptive written vocabulary levels?
(4) RQ4: What correlations exist between EMI students' proficiency (i.e., CEFR levels) and their mastery of receptive written vocabulary levels?

3. Methodology

3.1. Participants

Forty undergraduate students (13 males and 27 females) who speak Chinese as L1 (Cantonese or Mandarin) and were enrolled in a public university in Macau were recruited as participants in this study. Convenience sampling and snowball sampling (Cohen et al. 2011) were applied to recruit undergraduate students from different faculties (i.e., Arts and Humanities, Business Administration, Education, Law, Science and Technology, and Social Sciences), which resulted in students at Year 1 and Year 4 of undergraduate studies. The Year 1 undergraduate students had one semester of EMI learning experience, while the Year 4 students had been engaged in the EMI learning environment for seven consecutive semesters. Each participant was given the description of the study and an informed consent form. Demographic information about the participants was collected using an online questionnaire (see Appendix A), which included age (M = 19.975, SD = 1.717) and cumulative grade point average (CGPA) on a 4.0 scale (M = 3.125, SD = 0.502). The research methodology was reviewed and approved by the University of Macau Research Committee–Panel on Research Ethics.

3.2. Instruments

3.2.1. Listening Vocabulary Levels Test

The LVLT (Kramer n.d.; McLean et al. 2015) measures participants' English aural vocabulary knowledge from five 1000-word frequency levels and the Academic Word List (AWL). The unidimensionality of the LVLT was confirmed by McLean et al. (2015). The test comprises six sections with 150 items in total. The LVLT is a multiple-choice format test. The first five sections correspond to the five frequency levels of the most frequent 5000 word families (e.g., section one measures the most frequent 1000 word families, section two measures the most frequent 2000 word families), and each of these five sections contains 24 items. The AWL is included in the sixth section of the LVLT, containing 30 items that assess academic word knowledge. The original version of the LVLT was designed for

Japanese L1 speakers, in which the answer options were presented in Japanese. In this study, Chinese L1 speakers were invited to sit the test, so an L1 Chinese answer sheet was adopted for the present study. The Chinese version answer sheet was then further adapted so that the simplified Chinese version could be given to Mandarin speakers/readers and the traditional Chinese version could be given to Cantonese speakers. In the LVLT, participants heard a word followed by a straightforward carrier sentence that provided them with the grammatical context to access the meaning of the target word. Then, they had around five seconds to circle the answer that corresponds most closely to the meaning of the target word among four answer options. To determine participants' mastery level of aural vocabulary, a 21/24 cut-off score was used for the first 5000 frequency levels, and a 26/30 cut-off score was used for mastery of the AWL in accordance with Schmitt et al.'s (2001) suggested 26/30 criterion per level. According to Schmitt et al. (2001), learners were presumed to have mastered the higher frequency levels if they had mastered the lower frequency levels, and this guideline was also followed in this study. A participant's aural vocabulary mastery level was determined by the highest level at which a participant reached mastery.

3.2.2. Updated Vocabulary Levels Test

The UVLT is a revised and validated version of the Vocabulary Levels Test (VLT) (Nation 1983; Schmitt et al. 2001) that is used most widely to measure L2 learners' receptive vocabulary knowledge (Webb and Sasao 2013). The UVLT's unidimensionality has previously been confirmed (Webb et al. 2017). This updated test includes the measurement of the most frequent 5000 word families, dividing them into five frequency levels (Webb et al. 2017). For each frequency level, there are 10 clusters of three words, with 30 words for assessment in total. In each cluster, the participants are required to match the description to the corresponding word out of a list of 6 choices. The UVLT was scored following the guideline of Webb et al. (2017). Obtaining 29 items correct out of 30 for the 1000, 2000, and 3000 most frequent words was considered mastery, whereas attaining at least 24 out of 30 for the most frequent 4000 and 5000 levels was considered mastery of those levels. This study considered the highest level that a participant reached as his/her mastered vocabulary level (Ha 2022).

3.2.3. The Core Skills Test (CKT)

The Core Skills Test, one of the tests in the EnglishScore suite, is an internationally recognized assessment developed by the British Council and is delivered to participants through the British Council English Score application (British Council 2022). Overall, each of the test takers will get a total score up to 599 that represents their general English proficiency, as well as four individual scores for their proficiency in listening, reading, grammar, and vocabulary. While sitting the test, participants engaged in various question formats, including fill-in-the-blanks, multiple choice, and reordering paragraphs, all centered around everyday life topics. The scores of the CKT are in alignment with the CEFR framework, describing a learner's proficiency level on a range from beginner to advanced (i.e., A1 to C2). After test completion, a result page allows test takers to compare their obtained scores to other standardized language tests (See Table 2). The CKT provides accurate and trusted evaluations of English language skills for both young adult and adult learners from different language backgrounds (British Council 2022). In this study, the result of the CKT was used as a measurement of CEFR levels.

Table 2. The relationship between EnglishScore test scores, CEFR levels and other international standardized language tests.

CEFR	EnglishScore	TOEIC	TOEFL ITP	IELTS
Pre A1	0–99	0–119	-	-
A1	100–199	120–224	-	-
A2	200–299	225–549	337–459	-

Table 2. *Cont.*

CEFR	EnglishScore	TOEIC	TOEFL ITP	IELTS
B1	300–399	550–784	460–542	4.0–5.0
B2	400–499	785–944	543–626	5.5–6.5
C1	500–599	945–990	627–677	7.0–8.0

Note: "-" indicates there is no equivalent score within the particular CEFR level.

3.3. Procedures

The procedures are schematized in Figure 1. The data collection was conducted in two sessions. In the first session, participants were asked to take the UVLT and the LVLT. These two tests were administered in paper-and-pen format in a university library discussion room. After session 1, participants would take a 10 min break and proceed to session 2. In the second session, participants downloaded the British Council EnglishScore application and registered a test-taker account using their personal email address. Then, they took the CKT on their own devices. After they completed the test, they took another 5 min break and were invited to fill in an online questionnaire about their demographic information (i.e., age, gender, year of study) and academic performance (i.e., CGPA).

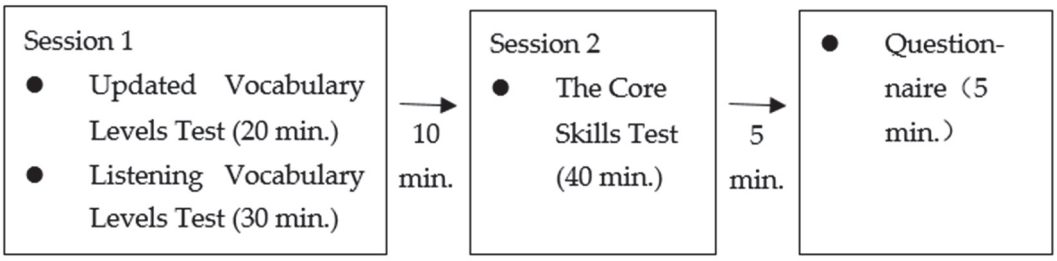

Figure 1. Research procedures.

3.4. Data Analysis

Descriptive statistics, correlation analysis, and regression analysis were conducted using SPSS 27.0. Descriptive statistics were computed to illustrate participants' mastery of English listening/reading vocabulary levels and CEFR levels. Moreover, correlation analyses were conducted to investigate the relationship between listening/reading vocabulary scores and CKT scores. Regression analysis was conducted to determine the extent to which the combination of receptive aural and written vocabulary knowledge predicts the CEFR level. Before conducting any inferential statistics, data analysis assumptions were tested. The data for CGPA, the CKT scores, the UVLT scores, the LVLT scores, and the AWL scores were not normally distributed, as shown by the results of Shapiro–Wilk tests: CGPA ($p = 0.018$), the CKT scores ($p = 0.008$), the UVLT scores ($p < 0.001$), the LVLT scores ($p = 0.004$), and the AWL scores ($p = 0.005$). Therefore, two-tailed non-parametric Spearman's rho correlations were conducted to investigate the relationship among CGPA, CKT scores, and receptive aural/written vocabulary scores (Larson-Hall 2010). Additionally, the coefficient alpha of the LVLT scores and the UVLT scores were calculated to ensure the reliability of the data. The Cronbach's alpha obtained was 0.669 for the UVLT scores and 0.696 for the LVLT scores, which indicated a fairly high level of internal consistency for vocabulary scores with this specific sample (Larson-Hall 2010). Before conducting the regression analysis, the assumptions for multiple regression were tested. However, the multicollinearity of the two variables (i.e., LVLT scores and UVLT scores) was not met ($r = 0.904$), indicating a high correlation between them. Additionally, the variance inflation factor (VIF) for each variable suggested that multicollinearity could be problematic (VIF = 5.449), as "values greater than 5 may harm the model" (Heiberger and Holland 2004, p. 243). To mitigate the multicollinearity issues, the dimensionality reduction technique of Principal Component Analysis (PCA) was applied to combine the correlated variables

into a single composite variable. This composite variable was then used in the regression analysis for this study.

4. Results

4.1. RQ1: To What Extent Is the English Proficiency of EMI Students (i.e., CEFR Levels) Associated with Their Mastery of Receptive Aural Vocabulary Levels?

The descriptive statistics of the mastery levels of the LVLT and language proficiency levels obtained from the EnglishScore test are presented in Table 3. The majority of participants clustered in CEFR B1 and B2 levels, within which participants exhibited varying levels of aural vocabulary mastery. Notably, the B2 level had the highest frequency of participants across all aural vocabulary mastery levels, with twelve participants having obtained an aural vocabulary mastery level of 2 k, six participants at the 3 k level, and two participants each at the 5 k and 1 k levels. Furthermore, participants at the CEFR C1 level displayed a diverse range of aural vocabulary mastery levels. One participant each was at the 2 k and 3 k levels, while two participants were at the 5 k level. Conversely, participants at the A1 and A2 CEFR levels had lower overall vocabulary mastery levels. However, exceptional cases were found at the A2 proficiency level, where two participants had an aural vocabulary mastery level of 3 k.

Table 3. Descriptive statistics of the CEFR and receptive aural vocabulary mastery levels.

		LVLT (*n* = 39)						**AWL (*n* = 40)**		
		1 k frequency level	2 k frequency level	3 k frequency level	4 k frequency level	5 k frequency level	Total	Mastered	Not mastered	Total
CEFR	A1	1	0	0	0	0	1	0	2	2
	A2	2	0	2	0	0	4	0	4	4
	B1	1	5	2	0	0	8	0	8	8
	B2	2	12	6	0	2	22	10	12	22
	C1	0	1	1	0	2	4	4	0	4
	Total	6	18	11	0	4	39	14	26	40

Note. C1 = advanced, B2 = upper intermediate, B1 = intermediate, A2 = elementary, A1 = beginner. One participant did not master any of the five frequency levels.

Therefore, participants classified in lower CEFR proficiency levels (i.e., A1 and A2) generally had lower aural vocabulary mastery levels (i.e., below 2 k level), while intermediate and advanced participants (i.e., B1 and above) had an overall higher aural vocabulary mastery level (i.e., above 2 k level). However, the varying aural vocabulary mastery levels within each CEFR proficiency level indicated the heterogeneity of participants' word knowledge. Therefore, a clear relationship between EMI students' CEFR levels and their aural receptive vocabulary levels was not identified in this study.

4.2. RQ2: What Correlations Exist between EMI Students' Language Proficiency (i.e., CEFR Levels) and Their Mastery of Receptive Aural Vocabulary Levels?

Table 4 displays the descriptive statistics, including measures such as means, ranges, standard deviations, and skewness for the five levels, the AWL, and the total scores of the LVLT. At the 1 k and 2 k frequency levels, the mean scores of the LVLT were greater than those at the other frequency levels and at the AWL. However, the mean score of the LVLT at the 5 k frequency level was slightly higher than that at the 4 k frequency level, which was similar to the results found in Wang's (2023) study. This was possibly because learners had encountered and learned the vocabulary at the 5 k frequency level in their past experiences (Wang 2023).

Table 4. Descriptive statistics of the LVLT and the UVLT test scores.

		Mean (%)	Maximum (%)	Minimum (%)	Range (%)	Standard Deviations (%)	Skewness
LVLT	1 k frequency level	98.7	100	83	17	3.2	−3.529
	2 k frequency level	90.9	100	46	54	10.6	−2.550
	3 k frequency level	78.0	100	38	62	15.1	−1.021
	4 k frequency level	67.2	100	29	71	15.6	−0.556
	5 k frequency level	69.9	100	38	62	13.5	−0.145
	Total score	80.8	99	51	48	9.8	−1.234
	The AWL	76.6	100	33	67	15.9	−0.981
UVLT	1 k frequency level	97.2	100	57	43	7.7	−4.202
	2 k frequency level	89.9	100	27	73	17.4	−2.355
	3 k frequency level	70.0	100	10	90	20.5	−0.898
	4 k frequency level	63.2	100	3	97	20.7	−0.754
	5 k frequency level	56.6	100	3	97	23.1	−0.467
	Total score	75.4	100	25	75	15.9	−1.483

To answer RQ2, correlation analyses were conducted among the participants' CKT scores, LVLT scores, AWL scores, and CGPA. Table 5 shows several statistically significant correlations between these variables. First, there was a positive correlation between participants' CKT scores and their LVLT scores (ρ = 0.624, p < 0.001). This indicates that participants with higher language proficiency tended to have a greater mastery of receptive aural vocabulary. Additionally, a strong positive correlation was also found between participants' CKT scores and their AWL scores (ρ = 0.654, p < 0.001), indicating that higher CKT scores were associated with higher AWL scores. Moreover, participants' mastery of receptive aural vocabulary was positively correlated to their mastery of the AWL (ρ = 0.806, p < 0.001). The result indicated that participants who achieved higher LVLT scores also performed well in the AWL section. Regarding the participants' academic performance, there was a positive correlation between EMI students' mastery of receptive aural vocabulary and their CGPA (ρ = 0.458, p = 0.003), suggesting that students with higher LVLT scores tended to achieve higher CGPA. Additionally, a statistically significant correlation was found between participants' AWL scores and their CGPA (ρ = 0.331, p = 0.037). However, the correlation between participants' CKT scores and their CGPA was relatively weak and not statistically significant (ρ = 0.289, p = 0.070), suggesting that there was no strong association between language proficiency and academic performance, as measured by CGPA.

Table 5. Correlations among CEFR scores, LVLT scores, LVLT scores, and CGPA.

	CKT Scores	LVLT Scores	AWL Scores	UVLT Scores	CGPA
CKT scores		0.624 **	0.654 **	0.682 **	0.289
LVLT scores	0.624 **		0.806 **	0.847 **	0.458 **
AWL scores	0.654 **	0.806 **		0.740 **	0.331 *
UVLT scores	0.682 **	0.847 **	0.740 **		0.401 *
CGPA	0.289	0.458 **	0.331 *	0.401 *	

Note. * p < 0.05, ** p < 0.01.

Therefore, the correlation analysis showed a positive correlation between EMI students' language proficiency and their mastery of receptive aural vocabulary. The statistically significant correlation suggested that as language proficiency increases, students are more likely to have a greater mastery of receptive aural vocabulary.

4.3. RQ3: To What Extent Is the English Proficiency of EMI Students (i.e., CEFR Levels) Associated with Their Mastery of Receptive Written Vocabulary Levels?

Table 6 shows the descriptive statistics of CEFR and receptive written vocabulary mastery levels, revealing that there was no clear relationship between the participants' CEFR

levels and their receptive written vocabulary mastery levels. This means that higher English proficiency does not necessarily equal broader vocabulary knowledge. For instance, only one C1-level participant mastered the most frequent 5000 word families, while six B2-level participants achieved this 5k level of mastery. Most participants were upper intermediate ($n = 22$), and few were beginners ($n = 2$). The number of advanced EFL learners equals that of the elementary level ($n = 4$).

There was variance in the UVLT scores of the participants ($M = 75.4\%$, $SD = 15.9\%$). Participants at each CEFR level demonstrated a diverse range of written vocabulary mastery levels, in which the widest range was found at the CEFR B2 level. Most participants ($n = 16$) were shown to have only mastered the most frequent 2000 word families, followed by the most frequent 1000 word families ($n = 8$), the most frequent 5000 word families ($n = 7$), and the most frequent 4000 word families ($n = 3$). The least number of participants ($n = 1$) were shown to master the most frequent 3000 word families. Five participants did not show mastery of any vocabulary levels.

Table 6. Descriptive statistics of CEFR and receptive written vocabulary mastery levels.

		UVLT (n = 35)					
		1 k frequency level	2 k frequency level	3 k frequency level	4 k frequency level	5 k frequency level	Total
CEFR	A1	0	0	0	0	0	0
	A2	0	1	0	0	0	1
	B1	3	5	0	0	0	8
	B2	5	7	1	3	6	22
	C1	0	3	0	0	1	4
	Total	8	16	1	3	7	35

Note. C1 = advanced, B2 = upper intermediate, B1 = intermediate, A2 = elementary, A1 = beginner. Five participants did not master any of the five frequency levels.

4.4. RQ4: What Correlations Exist between EMI Students' Language Proficiency (i.e., CEFR Levels) and Their Mastery of Receptive Written Vocabulary Levels?

Table 4 displays the descriptive statistics, including measures such as means, ranges, standard deviations, and skewness, for the five levels and the total scores of the UVLT. Among all the frequency levels, the mean scores of the UVLT at the 1 k and 2 k frequency levels were greater than those at the other frequency levels. Additionally, the mean score of each frequency level displayed a descending trend from 1 k to 5 k frequency levels.

To answer RQ4, correlation analyses were conducted among the participants' CKT scores, UVLT scores, and their CGPA. A statistically significant positive correlation was found between participants' CKT scores and UVLT scores ($\rho = 0.682$, $p < 0.001$). The statistically significant positive correlations suggested that higher English proficiency was associated with better receptive written vocabulary knowledge. Regarding the participants' academic performance, there was a positive correlation between EMI students' mastery of receptive written vocabulary and their CGPA ($\rho = 0.401$, $p = 0.010$), suggesting that students with higher UVLT scores tended to achieve higher CGPA.

4.5. Regression Analyses

Table 7 displays three regression models that significantly predict changes in the CEFR levels. After running the standardized multiple regression analyses, it was found that both LVLT scores and UVLT scores made significant contributions to participants' CEFR levels. Model 1, which included only LVLT scores, explains 58.3% of the variance in the CEFR levels. Model 2, which included only UVLT scores, explains 66.1% of the variance in the CEFR levels, slightly higher than the variance explained by Model 1 and Model 3. Model 3 used a composite variable of LVLT and UVLT scores, explaining 65.4% of the variance in CEFR levels. Therefore, Model 3 does not offer a significant advantage over Model 2.

Table 7. Standardized multiple regression models of LVLT scores and UVLT scores.

	Unstandardized Coefficients *B*	Standardized Error	Standardized Coefficient β	Adjusted R^2
Model 1				0.583 ***
Constant	−0.315	0.133		
LVLT scores	1.223	0.164	0.771	
Model 2				0.661 ***
Constant	0.068	0.070		
UVLT scores	0.802	0.091	0.818	
Model 3				0.654 ***
Constant	0.673	0.014		
Composite variable of LVLT scores and UVLT scores	0.127	0.015	0.814	

Note. *** $p < 0.001$.

5. Discussion

The findings suggest that while there was a general trend of higher aural/written vocabulary mastery levels among participants with higher CEFR levels, considerable variability within each CEFR level was observed. For example, there were participants at the CEFR B2 level with an aural vocabulary mastery level of 5 k and a written vocabulary master level of 2 k, whereas CEFR C1 participants only mastered the most frequent 2 k aural and written vocabulary. Previous studies showed that language in EFL textbooks was the major input for students' vocabulary learning in many EFL contexts (Alsaif and Milton 2012; Jordan and Gray 2019). Therefore, EFL learners' exposure to and engagement with English language materials might contribute to the observed differences in aural/written vocabulary mastery levels. Given the fact that L2 university students need to attain certain vocabulary knowledge to achieve 95% to 98% coverage of academic lectures (Dang 2022; Dang and Webb 2014; Reynolds et al. 2022, 2023), this finding highlighted the deficiency of the language requirement set for EMI university admission, which only examined applicants' English language proficiency indicated by CEFR levels and neglected their vocabulary mastery levels. Therefore, a connection between L2 learners' CEFR levels and their mastery of receptive aural/written vocabulary levels should ideally be established for EMI universities' reference in setting the language requirements for student admissions.

Although this study failed to map participants' receptive aural/written vocabulary mastery levels to their CEFR levels, results did indicate that greater receptive vocabulary knowledge was associated with higher CEFR levels, which is consistent with existing literature that highlights the positive correlation between vocabulary knowledge and language proficiency (Lin and Morrison 2010; Qian and Lin 2019; Rafique et al. 2023). While the findings of this study did not show a causative relationship between receptive vocabulary knowledge and CEFR levels, they suggested an intertwined and mutually reinforcing relationship between CEFR levels and vocabulary development (Zareva et al. 2005). This means that higher levels of vocabulary knowledge facilitates comprehension of the language, whereas higher CEFR levels provides a foundation for understanding and using newly learned vocabulary.

Furthermore, the positive correlation between receptive aural vocabulary knowledge and CGPA suggested that a solid aural vocabulary foundation is associated with better academic performance. Although this does not indicate a causation relationship between these two variables, the correlation may be explained by the EMI context in the university where lectures are delivered in English. Existing literature demonstrates a positive correlation between receptive aural vocabulary level mastery and L2 listening comprehension (Du et al. 2022; Matthews and Cheng 2015; McLean et al. 2015; Wallace 2020), which showed that a robust aural vocabulary foundation aids students' comprehension of EMI lectures, suggesting a strong predictive power of high frequency and academic words in L2 listening comprehension proficiency (Du et al. 2022). Therefore, establishing a link be-

tween L2 English learners' receptive aural vocabulary level mastery and their CEFR levels is crucial for ensuring the academic success of EMI university students.

Academic vocabulary knowledge played a crucial role in "the lengthy and difficult process of mastering the scale of lexicon needed to handle academic discourse" (p. 55) and was positively related to tertiary education achievement (Masrai and Milton 2018). The correlation between participants' AWL scores and CGPA was statistically significant in this study, which aligns with the findings of Masrai and Milton's (2018) study. While the study of Masrai and Milton (2018) showed the significance of receptive written academic vocabulary knowledge in L2 learners' academic success, the results of this study suggest the important role of receptive aural academic vocabulary knowledge in tertiary education achievement.

The regression analyses of this study showed that the UVLT might be a better predictor of students' CEFR levels compared to the LVLT. Students' CEFR levels were measured by the CKT, which was mainly comprised of written tasks. While there were listening tasks included in the CKT, they only accounted for one-third of the test. Therefore, the UVLT had a slightly stronger predictive power of CEFR levels than the LVLT.

Additionally, the correlation between the CKT scores and CGPA was not statistically significant. This finding was contradictory to that of most previous studies, where the linguistic knowledge of EMI university students was an important indicator of their successful performance in EMI studies in tertiary education settings (e.g., Chapple 2015; Hu et al. 2014). However, a study that investigated students' English proficiency and their EMI academic success in Turkey showed a similar result, finding a statistically non-significant correlation, which is explained by the limited impact of general English proficiency on students' EMI academic success compared to academic English proficiency (Curle et al. 2020). The CKT questions that participants in this study answered were mainly on daily topics, such as "going to the supermarket" and "conversations you may have in the office" (British Council n.d., p. 9), testing their general instead of academic English proficiency level. Therefore, the relationship between students' English proficiency and their EMI academic success was not statistically significant in this study, while it was statistically significant in studies that employed the measures of academic/content-related English proficiency, such as the final scores of English for Specific Purposes courses (e.g., Rose et al. 2020; Xie and Curle 2022).

Previous studies on EMI university students' English proficiency and their academic performance in the EMI setting recruited Year 2 participants from business programs in monolingual contexts such as Japan, China, and Turkey (e.g., Curle et al. 2020; Rose et al. 2020; Xie and Curle 2022). Participants of this study included Year 1 and Year 4 students from different faculties of a public EMI university in the multilingual context of Macao, providing a more comprehensive overview of L2 students' CEFR levels and their receptive vocabulary mastery levels.

6. Implications and Conclusions

In conclusion, this study did not find a clear pattern between L2 learners' receptive vocabulary mastery levels and their CEFR levels. However, statistically significant positive correlations were found between receptive aural/written vocabulary test scores and the CKT scores, indicating that higher receptive vocabulary test scores were associated with higher CKT scores. The results of this study provided implications for EMI universities on student admissions, classroom teaching, and the provision of additional English courses.

Firstly, EMI universities are encouraged to include an independent vocabulary level test in addition to an English proficiency test to evaluate their L2 English-speaking applicants. On one hand, the heterogeneity between L2 learners' receptive vocabulary mastery levels and their CEFR levels suggested that L2 learners who had met the minimum language requirement of an EMI university might encounter linguistic barriers in comprehending lectures. Multiple studies have also suggested that the mastery of the most frequent 2000 word families is insufficient for a comprehensive understanding of the EMI lec-

tures and learning materials (e.g., Reynolds et al. 2023; Dang and Webb 2014). On the other hand, the statistically significant positive correlation found between L2 learners' receptive aural vocabulary levels and their CGPA necessitates the inclusion of assessing L2 applicants' aural vocabulary level mastery as part of the admission process of EMI universities.

Secondly, the EMI lecturers are encouraged to make good use of EMI students' vocabulary profiles in selecting appropriate teaching materials and delivering lectures in a way that matches students' vocabulary levels to increase their comprehension of the content knowledge (Reynolds et al. 2023). The descriptive statistics of this study showed that higher English proficiency does not guarantee higher mastery of receptive vocabulary mastery levels. Moreover, different disciplines possess different demands of vocabulary knowledge on students (Reynolds et al. 2023). While giving lectures in lexically demanding fields, it is necessary for the EMI lecturers to notice and address the discrepancy between students' vocabulary levels and the demands of the discipline, which can be done by offering vocabulary lists for especially demanding lectures to students (Reynolds et al. 2023).

Lastly, the results of this study suggest refining the post-entry screening procedures to identify linguistically unprepared students and thus provide them with effective additional language courses. After being admitted to the EMI university, the L2 students are normally provided with various additional general and academic English courses to improve their readiness for studying in an EMI setting (Aizawa et al. 2023). Some of these courses are mandatory, while some of them are free electives, which depends on the students' proficiency levels shown in the university entrance exam or other internationally standardized English proficiency tests. For example, in a public university in Macau, students who are lower than the CEFR B1 level are required to take all the general and academic English courses; students who are in between the CEFR B1 and B2 levels only need to take academic English courses; students with CEFR B2 level or higher CEFR levels are waived from taking any of the English courses (University of Macau 2023). However, the unclear pattern between students' CEFR levels and their receptive aural/written vocabulary levels challenges the university's proficiency standard of waiving additional English courses for admitted students. Exempting CEFR B2 level students mastering 2000 word families or fewer of the most frequent 2000 word families could be problematic since these students are likely to encounter difficulties comprehending EMI lectures and instructional materials. Therefore, it is suggested to utilize students' vocabulary profiles in addition to their CEFR levels when conducting the screening process.

7. Limitations and Future Research

A key limitation of this study was the relatively small sample size, which may have contributed to the lack of a clear association between aural/written vocabulary mastery levels and CEFR proficiency levels. To obtain a more comprehensive understanding, future research should involve a larger group of EFL learners with diverse ages and proficiency levels. This would enable a deeper exploration of the specific factors that contribute to the observed heterogeneity in aural/written vocabulary mastery within CEFR proficiency levels, such as individual learner characteristics, instructional approaches, and exposure to authentic English language materials in EMI settings.

Additionally, the findings of this study may not be generalized to the entire EMI university student population or EFL learners of other age groups, as the participants were drawn only from university Year 1 and Year 4 students due to convenience sampling. However, this study has drawn researchers' attention to the discrepancy between students' receptive vocabulary mastery levels and the CEFR levels. These findings highlight the significance of further research in this area to support more effective language learning and academic success in EMI contexts. Although this study investigated the relationship between students' general English proficiency and their receptive aural/written vocabulary levels, future studies may consider using an academic English proficiency test to examine the correlations between these variables. In addition to receptive aural/written vocabulary

levels, future research may explore variables of productive aural/written vocabulary levels to strengthen the findings of this study.

Author Contributions: Conceptualization, B.L.R.; methodology, B.L.R.; software, Z.L. and J.Z.L.; validation, X.Z.; formal analysis, Z.L. and J.Z.L.; investigation, Z.L. and J.Z.L.; resources, B.L.R.; data curation, Z.L. and J.Z.L.; writing—original draft preparation, Z.L. and J.Z.L.; writing—review and editing, X.Z. and B.L.R.; supervision, B.L.R.; project administration, B.L.R.; funding acquisition, Z.L. and J.Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Macau, grant number HONR4001/2023-2/052/001.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the University of Macau Research Committee–Panel on Research Ethics, number SSHRE24-APP007-FED.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We would like to express our sincere gratitude to Paul Nation of Victoria University of Wellington for his valuable guidance and support during the preparation of this manuscript. Through personal communication, Nation generously provided us with his summary of vocabulary knowledge mapped to different CEFR levels. He also kindly clarified aspects of his notes when we had questions, helping to enhance our understanding.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Online Questionnaire

背景資料 Background Information 您在澳門大學就讀大學一年級/四年級嗎？ Are you a first-year/fourth-year student at the University of Macau?

O 是，我是大一學生 Yes, I am a Year 1 student
O 是，我是大四學生Yes, I am a Year 4 student
O 否 No

您的年齡是？ How old are you?

[ ]

您的性別是？ What is your gender?

O 男 Male
O 女 Female
O 不便告知 Prefer not to say

詳細資料 Detailed Information

您的學號是？ What is your student ID?

[ ]

您就讀的學院是？ What is your faculty?

O 人文學院 FAH
O 工商管理學院 FBA
O 教育學院 FED
O 健康科學學院 FHS
O 法學院 FLL
O 社會科學學院 FSS
O 科技學院 FST

您就讀的專業是？What is your major?

您在此學期取得的累計GPA成績是？What is your cumulative GPA until this semester of your study?

您就讀的中學的主要授課語言是？What language is used as the medium of instruction in your secondary school?

O 中文（粵語/普通話）Chinese (Cantonese/Mandarin)
O 葡語 Portuguese
O 英語 English
O 其他（請列明）Others (Please specify)

您在入學時提交的英語成績屬於以下哪個類別？What kind of language tests did you use for applying the university?

O 四校聯考 JAE Examination
O 高考 Gaokao
O 雅思 IELTS
O 托福 TOEFL
O 托福 iBT TOEFL iBT
O 多益 TOEIC
O 國際中學教育普通證書 IGCSE
O 其他（請列明）Others (Please specify the name of the examination and English score)

您在四校聯考中取得的英語成績為？What is your JAE Examination English score?

您在高考中取得的英語成績為？What is your Gaokao English score?

您在雅思考試中取得的英語成績為？What is your IELTS score?

您在托福考試中取得的英語成績為？What is your TOEFL score?

您在托福iBT考試中取得的英語成績為？What is your TOEFL iBT score?

您在多益中取得的英語成績為？What is your TOEIC score?

您在國際中學教育普通證書考試中取得的英語成績為？What is your IGCSE English score?

References

Aizawa, Ikuya, Heath Rose, Gene Thompson, and Samantha Curle. 2023. Beyond the threshold: Exploring English language proficiency, linguistic challenges, and academic language skills of Japanese students in an English medium instruction programme. *Language Teaching Research* 27: 837–61. [CrossRef]

Alsaif, Abdullah, and James Milton. 2012. Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal* 40: 21–33. [CrossRef]

Benigno, Veronica, and John de Jong. 2019. Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model. In *Developments in Language Education: A Memorial Volume in Honour of Sauli Takala*. Edited by Ari Huhta, Gudrun Erickson and Neus Figueras. Jyväskylä: University Printing House, pp. 8–29.

Bonk, William J. 2000. Second language lexical knowledge and listening comprehension. *International Journal of Listening* 14: 14–31. [CrossRef]

Botha, Werner. 2013. English-medium instruction at a university in Macau: Policy and realities. *World Englishes* 32: 461–75. [CrossRef]

British Council. 2022. Linking EnglishScore to the Common European Framework of Reference for Languages (CEFR). Available online: https://www.englishscore.com/resources/blog/how-does-englishscore-align-to-the-common-european-framework-of-reference-for-languages/ (accessed on 30 April 2024).

British Council. n.d. EnglishScore Core Skills Test Guide. Available online: https://es-mktg-assets.s3.eu-west-2.amazonaws.com/guide/englishscore-coreskills-guide-mobile.pdf (accessed on 30 April 2024).

Cai, Hongwen. 2020. Relating lexical and syntactic knowledge to academic English listening: The importance of construct representation. *Frontiers in Psychology* 11: 1–11. [CrossRef]

Chapple, Julian. 2015. Teaching in English is not necessarily the teaching of English. *International Education Studies* 8: 1–13. [CrossRef]

Cheng, Junyu, and Joshua Matthews. 2018. The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing* 35: 3–25. [CrossRef]

Cohen, Louis, Lawrence Manion, and Keith Morrison. 2011. *Research Methods in Education*, 7th ed. London: Routledge.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment–Companion Volume with New Descriptors, Council of Europe Publishing, Strasbourg. Available online: www.coe.int/lang-cefr (accessed on 30 April 2024).

Curle, Samantha Dogan Yuksel, Adem Soruc, and Mehmet Altay. 2020. Predictors of English medium instruction academic success: English proficiency versus first language medium. *System* 95: 1–29. [CrossRef]

Dang, Thi Nogc Yen. 2022. Vocabulary in academic lectures. *Journal of English for Academic Purposes* 58: 101123. [CrossRef]

Dang, Thi Ngoc Yen, and Stuart Webb. 2014. The lexical profile of academic spoken English. *English for Specific Purposes* 33: 66–76. [CrossRef]

Du, Guohui, Zuwati Hasim, and Fong Peng Chew. 2022. Contribution of English aural vocabulary size levels to L2 listening comprehension. *International Review of Applied Linguistics in Language Teaching* 60: 937–56. [CrossRef]

Enayat, Mostafa Janebi, and Ali Derakhshan. 2021. Vocabulary size and depth as predictors of second language speaking ability. *System* 99: 1–15.

ETS. 2010. Linking TOEFL iBT Scores to IELTS Scores: A Research Report. Available online: https://scholar.google.co.jp/scholar?hl=en&as_sdt=0,5&q=Linking+TOEFL+iBT+Scores+to+IELTS+Scores:+A+Research+Report&btnG=#d=gs_cit&t=1719713357688&u=/scholar?q=info:B77AC-ner7kJ:scholar.google.com/&output=cite&scirp=0&hl=en (accessed on 30 April 2024).

Field, John. 2008. *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Goh, Christine C. M. 2000. A cognitive perspective on language learners' listening comprehension problems. *System* 28: 55–75. [CrossRef]

Graves, Kathleen. 2008. The language curriculum: A social contextual perspective. *Language Teaching* 41: 147–81. [CrossRef]

Ha, Hung Tan. 2022. Issues in matching learners to their reading levels using receptive vocabulary knowledge–A closer look at the vocabulary levels tests and scoring methods. *Heliyon* 8: 1–7. [CrossRef]

Heiberger, Richard M., and Burt Holland. 2004. *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and SAS*. New York: Springer.

Hu, Guangwei, Linna Li, and Jun Lei. 2014. English-medium instruction at a Chinese university: Rhetoric and reality. *Language Policy* 13: 21–40. [CrossRef]

Hu, Jingjing, Xuesong (Andy) Gao, and Xuyan Qiu. 2021. Lexical coverage and readability of science textbooks for English-medium instruction secondary schools in Hong Kong. *Sage Open* 11: 1–9. [CrossRef]

Hulstijn, Jan H., J. Charles Alderson, and Rob Schoonen. 2010. Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. Edited by Inge Bartning, Maisr Martin and Ineke Vedder. Amsterdam: Eurosla, pp. 11–20.

Jordan, Geoff, and Humphrey Gray. 2019. We need to talk about coursebooks. *ELT Journal* 73: 438–46. [CrossRef]

Kramer, Brandon. n.d. *Vocabulary Tests*. Resources. Available online: https://www.brandonkramer.net/resources (accessed on 27 April 2024).

Larson-Hall, Jenifer. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. New York: Routledge.

Lin, Linda H. F., and Bruce Morrison. 2010. The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes* 9: 255–66. [CrossRef]

Macaro, E. 2022. English Medium Instruction: What do we know so far and what do we still need to find out? *Language Teaching* 55: 533–46. [CrossRef]

Masrai, Ahmed, and James Milton. 2018. Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes* 31: 44–57. [CrossRef]

Matthews, Joshua, and Junyu Cheng. 2015. Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System* 52: 1–13. [CrossRef]

McLean, Stuart, Brandon Kramer, and David Beglar. 2015. The creation and validation of a listening vocabulary levels test. *Language Teaching Research* 19: 741–60. [CrossRef]

Meara, Paul M., and James Milton. 2003. *The Swansea Levels Test*. Newbury: Express.

Milton, James. 2009. *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Milton, James. 2010. The development of vocabulary breadth across the CEFR levels. *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research* 1: 211–32.

Milton, James, and Thomai Alexiou. 2020. Vocabulary size assessment: Assessing the vocabulary needs of learners in relation to their CEFR goals. In *Vocabulary in Curriculum Planning*. Edited by Marina Dodigovic and Maria Pilar Agustín-Llach. Cham: Palgrave Macmillan, pp. 9–27.

Nation, Ian Stephen Paul. 1983. Testing and teaching vocabulary. *Guidelines* 5: 12–25.

Nguyen, Chi-Duc. 2021. Lexical features of reading passages in English-langauge textbooks for Vietnamese high-school students: Do hey foster both content and vocabulary gain? *RELC Journal* 52: 509–22. [CrossRef]

North, Brian, and Enrica Piccardo. 2019. Developing new CEFR descriptor scales and expanding the existing ones: Constructs, approaches and methodologies. *Zeitschrift für Fremdsprachenforschung* 30: 142–60.

Piccardo, Enrica. 2020. The Common European Framework of Reference (CEFR) in language education: Past, present, and future. *TIRF: Language Education in Review Series* 15: 1–13.

Qian, David D., and Linda H. F. Lin. 2019. The relationship between vocabulary knowledge and language proficiency. In *The Routledge Handbook of Vocabulary Studies*. London: Routledge, pp. 66–80.

Rafique, Samia, Ali Waqas, and Choudhry Shahid. 2023. The Correlation between Vocabulary Knowledge and English Language Proficiency at Undergraduate Level. *Pakistan Journal of Humanities and Social Sciences* 11: 1132–41. [CrossRef]

Reynolds, Barry Lee, and Melissa H. Yu. 2018. Addressing the language needs of administrative staff in Taiwan's internationalised higher education: Call for an English as a lingua franca curriculum to increase communicative competence and willingness to communicate. *Language and Education* 32: 147–66. [CrossRef]

Reynolds, Barry Lee, Xiaowen (Serina) Xie, and Quy Huynh Phu Pham. 2022. Incidental vocabulary acquisition from listening to English teacher education lectures: A case study from Macau higher education. *Frontiers in Psychology* 13: 1–18.

Reynolds, Barry Lee, Xiaowen (Serina) Xie, and Quy Huynh Phu Pham. 2023. The potentials for incidental vocabulary acquisition from listening to computer science academic lectures: A higher education corpus-based case study from Macau. *Frontiers in Psychology* 14: 1–13. [CrossRef] [PubMed]

Rose, Heath, Samantha Curle, Ikuya Aizawa, and Gene Thompson. 2020. What drives success in English medium taught courses? The interplay between language proficiency, academic skills, and motivation. *Studies in Higher Education* 45: 2149–61. [CrossRef]

Schmitt, Norbert, Diane Schimitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18: 55–88. [CrossRef]

Schoepp, Kevin. 2018. Predictive validity of the IELTS in an English as a medium of instruction environment. *Higher Education Quarterly* 72: 271–85. [CrossRef]

Siregar, Fenty L., and Henni Henni. 2023. An investigation of EFL students' vocabulary size and level at an Indonesian private tertiary institution. *Eralingua: Jurnal Pendidikan Bahasa Asing dan Sastra* 7: 305–15. [CrossRef]

Stæhr, Lars Stenius. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal* 36: 139–52. [CrossRef]

Stæhr, Lars Stenius. 2009. Vocabulary knowledge and advance listening comprehension in English as a foreign language. *Studies in Second Language Acquisition* 31: 577–607. [CrossRef]

Stoeckel, Tim, Stuart McLean, and Paul Nation. 2021. Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition* 43: 181–203. [CrossRef]

Taltz, Dietmar. 2011. English-medium masters' programmes at an Australian university of applied sciences: Attitudes, experiences and challenges. *Journal of English for Academic Purposes* 10: 252–70.

Uchihara, Takumi, and Tetsuo Harada. 2018. Roles of vocabulary knowledge for success in English-Medium Instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly* 52: 564–87. [CrossRef]

University of Macau. 2023. Course Requirements and Test out Information for "Language and Skills" Courses–English Language Courses. University of Macau. Available online: https://reg.um.edu.mo/download/ug-curriculum/TestOut-ENG-2023%E2%80%932024.pdf (accessed on 30 April 2024).

University of Macau. 2024. Application Requirement. University of Macau. Available online: https://reg.um.edu.mo/admissions/nonlocal/others-regions/app-adm-rules/ (accessed on 30 April 2024).

Wallace, Matthew P. 2020. Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning* 72: 5–44.

Wang, Yun. 2023. Explaining L2 learners' listening comprehension: The role of aural vocabulary knowledge. *Sintagma* 35: 103–18.

Wang, Yun, and Jeanine Treffers-Daller. 2017. Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System* 65: 139–150. [CrossRef]

Webb, Stuart A., and Yosuke Sasao. 2013. New directions in vocabulary testing. *Language Testing* 44: 263–77. [CrossRef]

Webb, Stuart, Yosuke Sasao, and Oliver Balance. 2017. The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics* 168: 33–69. [CrossRef]

Wu, Chiu-Hui. 2023. EMI as a problem, a right or a resource: Exploring misalignments in stakeholders' views at one university in Taiwan. *Language Teaching Research*. [CrossRef]

Xie, Wanying, and Samantha Curle. 2022. Success in English medium instruction in China: Significant indicators and implications. *International Journal of Bilingual Education and Bilingualism* 25: 585–97. [CrossRef]

Yang, Yingli, Ya Sun, Pengyun Chang, and Yumo Li. 2019. Exploring the relationship between language aptitude, vocabulary size, and EFL graduate students' L2 writing performance. *TESOL Quarterly* 53: 845–56. [CrossRef]

Yu, Shulin, Yan Wang, Lianjiang Jiang, and Bo Wang. 2020. Coping with EMI (English as a medium of instruction): Mainland China students' strategies at a university in Macau. *Innovations in Education and Teaching International* 58: 462–72. [CrossRef]

Zareva, Alla, Paula Schwanenflugel, and Yordanka Nikolova. 2005. Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition* 27: 567–95. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Exploring Lexical Bundles in the Move Structure of English Medical Research Abstracts: A Focus on Vocabulary Levels

Motoko Asano [1,*], Kensuke Hirosuna [2] and Miho Fujieda [1]

[1] Department of Foreign Languages, Faculty of Medicine, Osaka Medical and Pharmaceutical University, Takatsuki 569-8686, Japan; miho.fujieda@ompu.ac.jp
[2] Department of Regenerative Science, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama 700-8558, Japan; ken.twistedpair@gmail.com
* Correspondence: motoko.asano@ompu.ac.jp

Abstract: Research article abstracts, the second most-read part of research papers after titles, generally follow disciplinary conventions, which are often manifested in their language use. This study analyzed lexical bundles or multi-word sequences in move texts of a one-million-word corpus of English-language medical research article abstracts, with particular attention to vocabulary levels. The most frequent lexical bundles, such as "the primary end point was", often occurred once per text and predominantly took part in realizing a move. The coverage of the first thousand New General Service List was 63.6% for the entire corpus but was around 80% for bundles in Move 3, describing principal results, and those in Move 4, evaluating the results. Many of the sequences were research-oriented bundles, used to express research contexts. The bundles were made up of relatively accessible word items, but the sequences occurred to realize highly specific research contexts. The findings suggest that becoming familiar with the bundle may need increasing awareness of disciplinary conventions such as guideline adherences and statistical procedures. This study may offer insights on the need for learners to familiarize themselves with these bundles.

Keywords: English for academic purposes; lexical bundles; vocabulary; move; medical research article abstracts; corpus linguistics

1. Introduction

In an era where information floods every aspect of academic and professional life, understanding the conventions of academic writing genres, including textual and vocabulary features, is essential (Swales 1990; Bhatia 2019; Coxhead 2020). Over the past 50 years, the academic landscape has undergone unprecedented changes, marked by a substantial increase in the number of academic papers (Hyland and Jiang 2019). This rapid growth leads to a need for learners, especially medical students, to become proficient in reading and writing research article abstracts because of the specialized nature of their disciplinary texts (Coxhead 2016; Dang 2020; Simpson 2022; Tang and Liu 2019).

Functioning as an academic, especially in English, entails engaging in a sophisticated activity framework of communication practices, particularly when constructing and disseminating scientific claims (Belcher 2016). In this system, language plays a pivotal role in forging consensus and persuading the target discourse community, a group of individuals with shared goals (Swales 1990). Academic discourse is navigated through genre texts, which are types of spoken or written texts used for specific communicative purposes (Hyon 2018, p. 3).

Written communication can be distinguished by differences in audience, purposes, content, form, style, and context (Robinson et al. 2008). Academic written language is known to be differ from academic spoken language, particularly from a vocabulary perspective (Dang et al. 2017; Coxhead and Dang 2019). Students, particularly those

learning English as an additional language (EAL) need to learn the genres and conventions commonly used by community members (Brooks et al. 2023; Samraj 2016; Flowerdew 2000). This mastery is the focus of English for academic purposes (EAP) research. EAP research has two main pillars: analyzing genre-specific texts to identify communicative functions and structures (Samraj 2016), and vocabulary studies (Coxhead 2016). Both aim to aid learners' academic development (Hyland 2016).

EAP research often involves constructing corpora to facilitate these analyses (Handford 2010). One of the central pillars of this research is genre analysis (Swales 1990). The other main pillar of EAP research, vocabulary studies, also employs corpus linguistics techniques (Nation 2005). As Storch et al. (2016) indicated, approaches to EAP are influenced by various factors, including differences in higher education systems across countries. Therefore, research-based instruction integrating multiple approaches is recommended (Hyland 2016; Storch et al. 2016). In this context, linking vocabulary quantification with rhetorical moves has been attempted by several studies (Cortes 2013; Mizumoto et al. 2017; Qi and Pan 2020; Casal and Kessler 2020). These studies illustrate the importance of combining move analysis and vocabulary studies to shed light on the features of disciplinary texts. However, to our knowledge, studies examining the vocabulary levels of word items in lexical bundles within the moves of abstracts have been scarce. Therefore, this study aimed to examine lexical bundles across moves in disciplinary research article abstracts with a focus on the vocabulary levels.

2. Literature Review

2.1. Move Analysis

As a key element of genre analysis framework, move analysis is "a text analytical method developed by Swales in 1981" (Moreno and Swales 2018, p. 40). Moves, as defined by Swales (2004, pp. 228–29), are "discoursal or rhetorical units performing coherent communicative functions in texts", exhibiting variability in length and other features. The "revised *Create a Research Space* (CARS) model" (Swales 1990, p. 140), preceded by "a 4-Move Schema" (Swales 1981, p. 15), provides an analytical approach to understanding the structure of research article introductions. Swales (2004) refined the CARS model to include distinct move structures: "Move 1 Establishing a territory", "Move 2 Establishing a niche", and "Move 3 Presenting the present work" (pp. 230–32). These moves, denoting their communicative purposes, can be segmented into "Steps" (Swales 2004, p. 230) that realize "functional components" (Tankó 2017, p. 43). The move analysis approach has been applied extensively to examine research articles as a whole (Kanoksilapatham 2005; Maswana et al. 2015; Mizumoto et al. 2017; Nwogu 1997; Stoller and Robinson 2013) and abstracts (Hyland 2000; Kanoksilapatham 2009; Pho 2008; Salager-Meyer 1991, 1992; Tankó 2017) in various disciplines, including medicine.

2.2. Vocabulary Studies

Teaching common words in specific contexts helps learners effectively improve their vocabulary (Nation 2001). In support of learners, various word lists have been developed from collections of specialized texts (Browne 2013; Coxhead 2000; Coxhead and Hirsh 2007; Fraser 2007; Tang and Liu 2019; Wang et al. 2008). An early example is Thorndike's word book for instructors (Thorndike 1921), which was later updated by Thorndike and Lorge (1944). The General Service List (GSL; West 1953) was subsequently compiled after extensive international discussions before and after World War II (Gilner 2011). Developed from a five-million-word corpus, the GSL is tailored for ESL/EFL learners and is organized into two groups of word families (Coxhead 2000). The GSL has frequently been used as a basis for developing other word lists, including the Academic Word List (AWL) of 570 word families prepared by Coxhead (2000).

Many studies utilize the GSL and AWL to develop specialized word lists. Coxhead and Hirsh (2007) created "the pilot science corpus" (p. 70) from textbooks and discipline-specific reading materials to compile a word list covering items not included in the GSL

or AWL. Fraser (2007) quantified the vocabulary required by pharmacology students and compiled the Pharmacology Word List (PWL) from international pharmacology journal articles. Wang et al. (2008) investigated key medical vocabulary to support curriculum design and learning objectives and created a Medical Academic Word List (MAWL). To support medical students, Quero and Coxhead (2018) also compiled word type lists based on their medical corpora, which included texts from medical textbooks. These studies are just a few examples of the many that have used the GSL and AWL to examine the features of their corpora.

While the original GSL has been a foundational tool in language learning over 50 years, it has faced some criticism in recent time (Green and Lambert 2018). The approach of utilizing word families in the GSL and AWL has been the subject of some debate among researchers (Gardner and Davies 2014). In response to these discussions, the New General Service List (NGSL), developed by Browne and his colleagues (Browne 2013), was designed as a modern update of the original GSL. Consisting of 2801 high-frequency words with a clearer definition of a "word" and a higher coverage of general English (Dang and Webb 2016), the NGSL is viewed by some researchers as more suitable for EAL learners (Mizumoto et al. 2021). It has been suggested that EAL undergraduates may find the NGSL to offer a slightly easier learning curve, potentially making it more appropriate for learners compared to the GSL (Culligan 2019). Subsequently, the New Academic Word List (NAWL), consisting of 963 words, was developed to align with the NGSL along with other specific purpose word lists targeting mid-frequency range vocabulary. It is reported that the NGSL and the NAWL combined offer an average of 92% coverage of academic texts and lectures (Browne 2021, p. 4). These updates reflect ongoing efforts to improve vocabulary learning tools for English language learners.

2.3. Lexical Bundles

In vocabulary studies that support learning academic texts, research on multiword units, often referred to as "*lexical bundles*", is also significant (Nation 2022, p. 454). Corpus-based studies have identified these lexical bundles (Samraj 2016). The term lexical bundles is defined as "sequences of word forms" (Biber et al. 1999, p. 990) that occur frequently across texts, characterizing a genre or discipline. This concept has gained prominence in corpus linguistics research (Biber et al. 1999; Cortes 2004, 2013; Hyland 2008a). Lexical bundles have been studied extensively for helping learners develop a repertoire that is sensitive to disciplinary norms (Hyland 2012, p. 17).

A study by Cortes (2013) associated the identification of lexical bundles and rhetorical analysis of moves and steps in research article introductions. This approach has contributed to revealing essential relationships between forms and functions (Gray et al. 2020, p. 139), showing the values of prefabricated word sequences (Biber et al. 2004, p. 376) within the linguistic contexts that constitute "communicative events" (Swales 1990, p. 9).

Following Cortes (2013), several scholars have observed move-specific bundles (Mizumoto et al. 2017; Omidian et al. 2018; Qi and Pan 2020). Mizumoto et al. (2017) observed 25 moves across 1000 research articles as a whole in applied linguistics, identifying frequently occurring lexical bundles in each move. Omidian et al. (2018) identified lexical bundles in each move in research article abstracts from six disciplines such as mechanical engineering, physics, and applied linguistics. Their study has shown a marked difference in the frequency of bundles across disciplines, suggesting that researchers from different fields prioritize different aspects when presenting their work in academic abstracts (Omidian et al. 2018, p. 12). Qi and Pan (2020) performed move analysis according to Hanidar's (2016) four-move structure and extracted lexical bundles, suggesting the role of lexical bundles in achieving the communicative goals of rhetorical sections.

These bundles, also known as "multi-word sequences" (Biber et al. 2004, p. 373; Mizumoto 2015, p. 30) or "recurrent word combination" (Chen and Baker 2010, p. 31), and synonymous with "n-grams" (Mizumoto 2015, p. 31; Stubbs and Barth 2003, p. 61), play a crucial role in the structure of discourse. These bundles often encapsulate shorter sequences,

such as three-word bundles, within their four-word structures (Cortes 2004, p. 401; Hyland 2008a, p. 6). Although Stubbs and Barth (2003) question the status of n-grams as linguistic units and refer to them as "chains of word-forms" (p. 62), they acknowledge that these recurring sequences typify specific text genres. Traditionally, studies have focused on the prevalence of four-grams in academic writing, with phrases like "as a result of" and "in the context of" serving as organizers of specific texts (Cortes 2013, p. 34).

Hyland (2008a, p. 6) has shown that frequency counts "drop dramatically" when quantitating five-grams compared to four-grams, stating that "many four and five word strings" share three-grams. Cortes (2024, pp. 120–21) discusses "potential overlaps" such as the five-gram "as a result of the" extending beyond shorter sequences like "as a result". Furthermore, Hyland and Jiang (2019, p. 110) have demonstrated that longer strings, such as the five-gram "at the beginning of the" which extend beyond the four-gram "the beginning of the" may offer deeper insights into the pedagogical applications of lexical bundles. Golparvar and Barabadi (2020, p. 6) analyzed "phrase frames" (p-frames; Cortes 2024, p. 105) in the discussion section of research articles and suggested that five-word p-frames could often be "more specific to a particular genre" compared to four-word p-frames. The most frequently occurring five-word p-frame was "are * likely to withdraw", with fillers such as "less" and "more", while the leading four-word p-frame was "the * of technology", with fillers such as "use" and "adoption". Liu and Chen (2022) have shown that five-word and six-word p-frames in university lectures reveal knowledge-disseminating and content-oriented features. Casal and Kessler (2020), who studied frequent p-frames in grant applications, demonstrated a strong association between the writers' use of five-word p-frames and their rhetorical intentions. Using a bundle-driven approach, Li et al. (2020) found that five-word bundles contributed to identifying moves of PhD abstracts. These findings indicate that extended sequences such as five-grams may be useful for genre-specific language use and potentially provide linguistic patterns beneficial for learners.

2.4. Medical Research Abstracts

Abstracts are the most-read texts in a research paper, second only to titles, and serve as a vital standalone tool of communication (Hyland 2000). Scientists increasingly rely on abstracts as "short, concise, complete, and accurate sources of information" (Salager-Meyer et al. 2014, p. 222). Research paper abstracts act as "advance indicators of the content and structure of the following text" (Swales 1990, p. 179) and encapsulate the key findings of the research (Huckin and Olsen 1983, p. 359).

Structured abstracts have become the norm in medical literature, largely because of long-standing endorsements by journal editors, such as those from the International Committee of Medical Journal Editors (ICMJE) since 1993 (Salager-Meyer et al. 2014, p. 223). Salager-Meyer (1991) noted that among 77 medical research article abstracts, half were deemed "poorly structured" in adherence to "the Introduction, the Methods, the Results, and the Discussion (IMRAD) pattern" (p. 528). While Anderson and Maclean (1997) examined "unstructured abstracts" (p. 2) of medical article abstracts and identified a five-move structure, including an optional background, along with the purpose, method, results, and conclusion, Hanidar (2016), analyzing abstracts from various disciplines including medicine, advocated for a four-move structure, comprising "Move 1: Creating a research space, " "Move 2: Describing research procedure", "Move 3: Summarizing principal results", and "Move 4: Evaluating results". This four-move approach aligns with our corpus texts structured according to the journal's guidelines. The studies of structured abstracts, notably by James Hartley, highlights the benefits of structured texts over traditional formats (Hartley 1993, p. 90). Hartley (1999, 2003) argued that structured abstracts, already used in medicine, could be "appropriate for applied ergonomics" (1999, p. 535) and could be "introduced into psychology journals" (2003, p. 366). His extensive work on structured abstracts was comprehensively reviewed by Zhang and Liu (2011), who have underscored "the advantages of structured abstracts over traditional ones" (p. 575).

In medical research publications, articles are often categorized by study design and must meet the requirements of specific reporting standards (Millar et al. 2019; Stosic 2022). High-quality reporting is facilitated by over 616 standardized reporting guidelines from the EQUATOR Network (2024), up from around 400 different sets (Millar et al. 2019, p. 150). Failure to adopt the guidelines may cause a manuscript to be regarded as inferior in quality by the ICMJE (Millar et al. 2019, p. 141). The ICMJE (2024) aims to enhance the quality and transparency of medical reporting. Their recommendations are broadly "accepted by biomedical journals" (Luo and Hyland 2019, p. 39) and play a crucial role (Millar et al. 2012, p. 393) in shaping the research writing standards. These initiatives appear to accelerate the standardization of research writing (Swales 2017, p. 249).

While there are no specific guidelines on language use (Stosic 2022), medical abstract conventions have been studied from the perspective of linguistic features. Salager-Meyer (1992) analyzed 84 medical abstracts and revealed the choice of verb tense and modality associated with rhetorical functions. Abdollahpour and Gholami (2018) gathered 1800 medical article abstracts from various journals. They identified four-word lexical bundles and classified them into general and technical groups by two qualified raters (Abdollahpour and Gholami 2018). Nam et al. (2016), who aimed to reformat unstructured abstracts into the IMRAD format, have found that linguistic features such as a five-word sequence "aim of this study was" would improve the effectiveness of abstract sentence classification. These studies have contributed to understanding the textual features of medical research article abstracts.

2.5. Application of Move Analysis and Vocabulary Research

Most recent studies have integrated move analysis and vocabulary research to envisage academic writing instruction (Casal and Kessler 2020; Li et al. 2020; Qi and Pan 2020). As a practical application, Mizumoto et al. (2017) developed an academic writing support tool called the "Academic Word Suggestion Machine (AWSuM)". AWSuM integrates move analysis with lexical bundle analysis. John Morley's (2023) "The Academic Phrasebank" serves as a general resource for academic writers. It offers examples of phraseological components organized according to the main sections of a research paper or dissertation. This integrated approach is particularly beneficial for EAL learners. For instance, first-year medical students are shown to have difficulties in reading medical research abstracts (Shimizu 2019). It was pointed out that the boundary of methods and results sections is hard to identify when the learners face sentences like "*Of the 19,114 persons who were enrolled in the trial, 9525 were assigned to receive aspirin and 9589 to receive placebo*" (Shimizu 2019, p. 85). These findings echo Tardy and Jwa's (2016) observations about the challenges of learning academic writing within a discipline, suggesting the importance of integrating genre and vocabulary research for future instructional practices.

3. The Current Study

The purpose of this study was to investigate the vocabulary levels of medical research article abstracts, especially lexical bundles across moves in disciplinary research article abstracts. This study identified lexical bundles across moves of research article abstracts and used the New General Service List (NGSL; Browne 2013) for examining the vocabulary levels of word items in the move texts and lexical bundles across moves. Our study poses the following research questions:

Research Question 1: Which lexical bundles occur most frequently in specific moves within medical research abstracts?
Research Question 2: What are the language features of lexical bundles across moves in medical research abstracts, such as the coverage of the NGSL, forms, and functions?

4. Data and Methods

4.1. Corpus

The data used in this study contain abstracts from the original research articles published in the years 2002–2020 in *The New England Journal of Medicine* (Table 1). These abstracts were selected for their "representativity, reputation, and accessibility", according to Nwogu's (1997, p. 121) criteria. The journal's articles are also valued for providing regional students with essential insights into evaluating medical literature (Ogawa 2014) and understanding the practical use of language in medical writing (Jego 2012). Additionally, the journal provides official translations in the local language (Nankodo 2024).[1]

Table 1. Corpus of research article abstracts.

Description	Data
Years	2002–2020
Type of texts	Research article abstracts
Number of texts	3983
Mean length of texts	303
Total number of words (tokens)	1,148,583
Total word types	22,950
NGSL and NAWL coverage	
First NGSL	63.6%
First and second NGSL	71.1%
First, second, and third NGSL	75.6%
First, second, third NGSL, and NAWL	80.7%
Move tokens	
Move 1	162,558
Move 2	317,349
Move 3	497,215
Move 4	171,461
Move word types	
Move 1	11,897
Move 2	13,631
Move 3	14,229
Move 4	12,169

Each article was segmented into individual sentences. Drawing on the studies by Cortes (2013, p. 36), who identified lexical bundles in research article introductions and mapped them to specific rhetorical moves, and Mizumoto et al. (2017, p. 902), who concentrated on extracting "move-specific lexical bundles", we segmented our corpus by sections (moves). For move identification, we adopted the methodology of Qi and Pan (2020), dividing our corpus texts "based on their own headings" (p. 112)—Background, Methods, Results, and Conclusions—to create subcorpora "Move 1: Creating a research space", "Move 2: Describing research procedure", "Move 3: Summarizing principal results", and "Move 4: Evaluating results" according to Hanidar (2016, p. 14). This four-move approach aligns with our corpus texts structured according to the journal's guidelines (The New England Journal of Medicine 2024).

The abstract texts had 1,148,583 words, as determined using AntConc (Version 4.2.4; Anthony 2023).[2] The number of words in each text was quantitated using CasualConc (Version 3.0.8; Imao 2024) and was visualized using Google Colaboratory's Python environment (Version 3.10.12). The average word count of the abstracts was 303 words, with a standard deviation of 44 words. The distribution of abstract word counts showed that 50% of the texts fell between 273 and 332 words, with a median value of 299 words (Figure 1). These findings indicate a relatively consistent length across the corpus texts.

The most frequent words in the medical research abstract corpus included many function words, with *the* and *of* ranked the first and second (Table 2). The top 100 words accounted for over 51% of the texts (Table 3). These findings suggest that a limited set of words were quite frequently used for presenting research information. Furthermore,

frequent instances of these function words and the scarcity of specialized terms suggest that studies on diverse topics were reported using similar basic vocabulary.

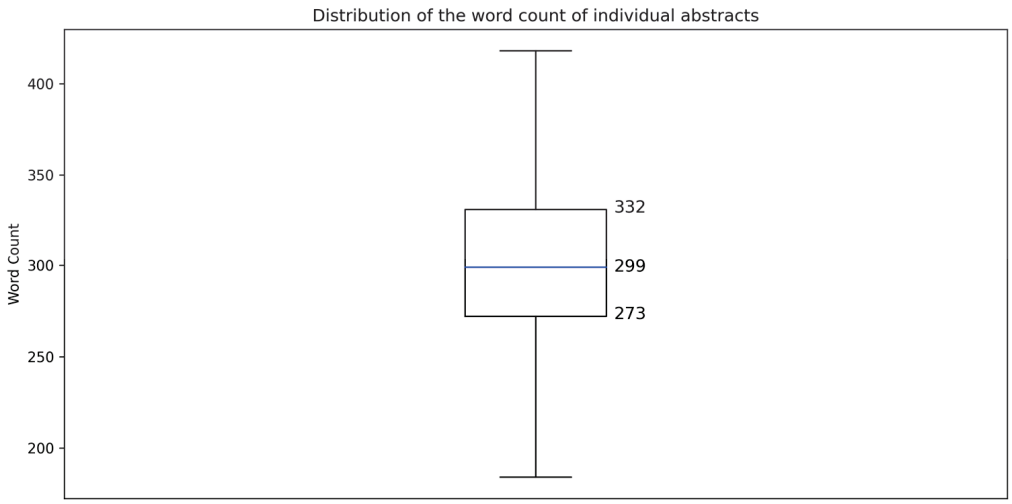

Figure 1. Distribution of the word count of individual abstracts.

Table 2. List of the most frequent words in the abstract corpus.

Rank	Word	Freq per Mil *	Rank	Word	Freq per Mil *
1	the	52,875	26	that	4026
2	of	46,186	27	among	3982
3	in	33,594	28	than	3944
4	and	29,721	29	placebo	3852
5	to	22,601	30	an	3674
6	with	20,870	31	treatment	3428
7	a	17,658	32	after	3372
8	patients	14,357	33	is	3192
9	was	11,982	34	on	3148
10	for	10,868	35	years	3135
11	were	10,866	36	disease	3131
12	or	9611	37	not	3042
13	group	9493	38	ratio	3001
14	p	6739	39	therapy	2936
15	at	6410	40	primary	2772
16	we	5846	41	mg	2721
17	by	5437	42	associated	2714
18	from	5082	43	rate	2699
19	had	4913	44	interval	2619
20	who	4801	45	months	2585
21	as	4451	46	death	2581
22	ci	4073	47	confidence	2492
23	percent	4038	48	number	2470
24	risk	4031	49	assigned	2468
25	per	4030	50	compared	2420

* Freq per mil stands for frequency per million words.

Table 3. Cumulative coverage of the top 100 words in the abstract corpus.

Rank	Coverage	Rank	Coverage
10	26.1%	60	44.8%
20	33.0%	70	46.7%
30	37.0%	80	48.4%
40	40.1%	90	49.9%
50	42.7%	100	51.4%

For the analysis with the NGSL (Version 1.01, Browne 2013) and NAWL (Version 1.01, Browne et al. 2013), we used a corpus tool, CasualConc (Imao 2024). The NGSL comprises a total of 2801 words. Separate lists categorizing the top 1000 words by frequency (first NGSL), the next 1000 words (second NGSL), and the remaining 801 words (third NGSL) were available from the AntWordProfiler website (Anthony 2024) as a resource for vocabulary profiling. In addition, a supplementary list of 174 basic words, which were not included in the aforementioned three lists, and the NAWL, consisting of 963 words, were also downloaded from the same site. By combining these five lists cumulatively, the following four stopword lists were created and imported into CasualConc for coverage analysis.

1. First NGSL + Supplement
2. First NGSL + Second NGSL + Supplement
3. First NGSL + Second NGSL + Third NGSL + Supplement
4. First NGSL + Second NGSL + Third NGSL + NAWL + Supplement

Using the stopword function of CasualConc, the word items in the lists were applied to automatically remove the target items in the preparatory step (Sarica and Luo 2021). To determine the covered word count, the word count after removal was subtracted from the total word count of the texts using Microsoft Excel (Version 2406). This procedure was repeated using the aforementioned stopword lists one by one, and the results were visualized using Python 3.10.12 in the Google Colaboratory environment (Figure 2).

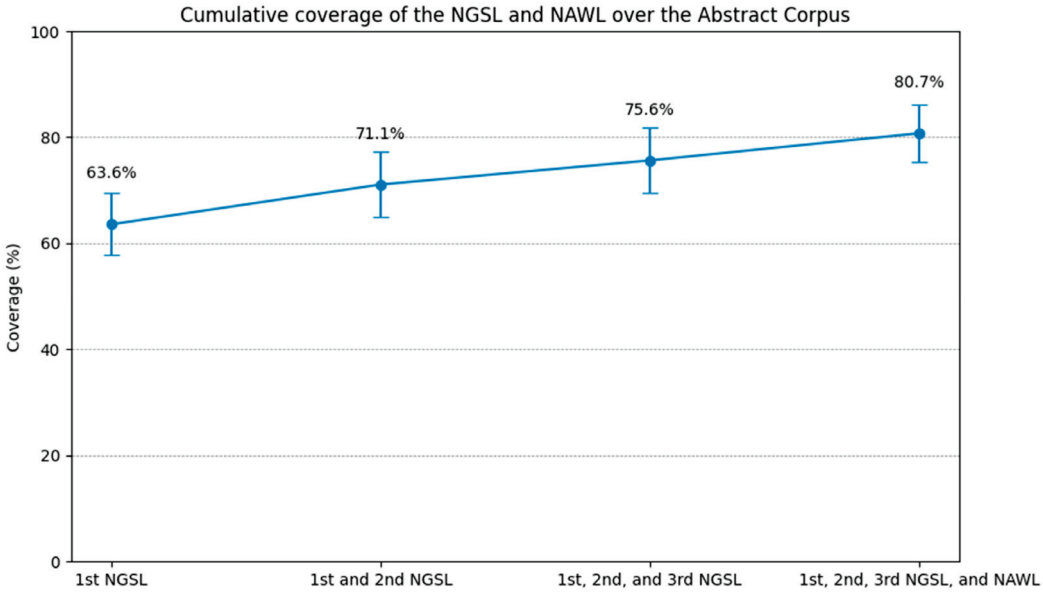

Figure 2. Cumulative coverage of the NGSL and NAWL over the abstract corpus.

The first 1000 words plus the supplementary 172 words covered 63.6% of the entire corpus, 71.1% with the first 2000 words, and 75.6% with the addition of the remaining 801 words. The NAWL yielded an additional 5.1% coverage, reaching a total coverage of 80.7%.

The comparison with previous studies showed that while the NGSL and NAWL provided greater coverage in other texts, they did not cover as much of our medical research article abstract corpus. The 2801 high-frequency words from the NGSL provided between 95% and 97% overall text coverage of English reading passages of Japan's national university entrance examination from 2015 to 2019 (MacDonald 2019, p. 22). The NGSL and NAWL combined to provide an average coverage of 84.8% of a 7.8-million-word civil engineering research article corpus (Gilmore and Millar 2018). In contrast, the coverage of our medical research abstract corpus was lower, indicating its highly specialized vocabulary level.

4.2. Data Processing

Hyland and Jiang (2019, p. 111), following Cortes (2015, p. 205), argued that while bundle frequencies are often standardized per 10,000 words, this normalization may result in higher instances of bundles in smaller corpora. This potentially raises the frequency of word combinations that are not usually common enough to surpass the threshold. Consequently, phrases that are infrequently used might still be classified as lexical bundles after normalization (Hyland and Jiang 2019, p. 111). To mitigate this issue, we adopted multi-step approach to bundle identification, drawing on methodologies mainly from Cortes (2004), Hyland (2008a), Hyland and Jiang (2019), and Lake and Cortes (2020). Our initial step was to ensure that each abstract in our corpus had broadly similar word count in English and quantitated the abstracts (Figure 1).

Recent studies have shown that five-word bundles are likely to realize rhetorical intentions (Casal and Kessler 2020) and represent a particular genre (Golparvar and Barabadi 2020). Nam et al. (2016) studied several linguistic features for classifying unstructured abstract sentences into the IMRAD format. They found that n-grams produced the best results. However, increasing the value of "n" in n-grams did not necessarily improve classification performance. Better results were obtained with sequences such as five-grams and six-grams (Nam et al. 2016). Therefore, the present study focused on identifying five-word bundles. Initially, following the criteria for five-word p-frames set by Casal and Kessler (2020), five-word bundles appearing in at least five different texts with a raw frequency of five were identified. This was achieved with the N-Gram function of AntConc (Version 4.2.4, Anthony 2023). This process revealed 433 bundles in Move 1, 1901 in Move 2, 3349 in Move 3, and 840 in Move 4. Given the size of the move corpora (Table 1), we followed the idea of Bestgen (2020) to avoid the potential bias of smaller corpora yielding more bundles by using higher frequency thresholds. For further analysis, a threshold of appearing in at least five different texts with a raw frequency of 10 was applied.

We followed the approach of Hyland and Jiang (2019, p. 109), which involved manually removing bundles containing text-dependent noun phrases, such as "the United States". However, we retained phrases such as "the primary end point". This is because primary end points serve as key measures of study outcomes, as noted by Qi and Pan (2020, p. 115), and because of the substantial instructional value of items conforming to disciplinary requirements (Salager-Meyer et al. 2014; Luo and Hyland 2019; Millar et al. 2019). This strategy aimed at extracting bundles while preserving "the observation of language in use" (Sinclair 1991, p. 39) as much as possible. In this process, we followed Durrant's (2017, p. 170) study for the exclusion of "punctuation and numerals". We utilized AntConc's feature for automatically omitting punctuation and digits (Viana and O'Boyle 2022, p. 119). Then, we manually excluded "overlapping word sequences" (Chen and Baker 2010, p. 33) with reference to the procedure outlined by Qi and Pan (2020, p. 113). This was done by placing the extracted bundles on spreadsheets (Microsoft Excel) and examining AntConc's concordance lines. For example, the bundle "did not reduce the rate" showed 13 times in Move 4; all instances were from the six-word string "did not reduce the rate of". The five-word string "not reduce the rate of" occurred 15 times, including two instances of a six-word sequence, "does not reduce the rate of" (Figure 3). Therefore, the bundle "not reduce the rate of" was listed.

The structures of the bundles were examined with reference to Biber et al. (1999, pp. 1014–24). We also referred to the procedure by Hyland (2008a, p. 10) and examples by Qi and Pan (2020, pp. 125–28). Biber et al. (1999) presented taxonomies for lexical bundles in academic prose, providing example bundles embedded in sentences. Hyland (2008a), citing Biber et al. (1999), also gave example bundles in sentences. By comparing our identified bundles to those examples, we classified our bundles accordingly. Bundles that did not fall into the categories based on the examples, such as "than in the group that", were labeled as others.

The functions of the bundles were analyzed based on Hyland's (2008a) classification. Hyland (2008a, p. 13) explained that research-oriented bundles "help writers to structure

their activities and experiences of the real world". Accordingly, our observation included both less specific items such as "with a use of the" and highly specific combinations such as "hazard ratio confidence interval ci" in this category. Text-oriented bundles are "concerned with the organisation of the text and its meaning as a message or argument", and participant-oriented bundles are "focused on the writer or reader of the text" (Hyland 2008a, pp. 13–14). Based on these definitions, bundles such as "associated with an increase in" were differentiated from research-oriented sequences such as "associated with an increased risk" in classification.

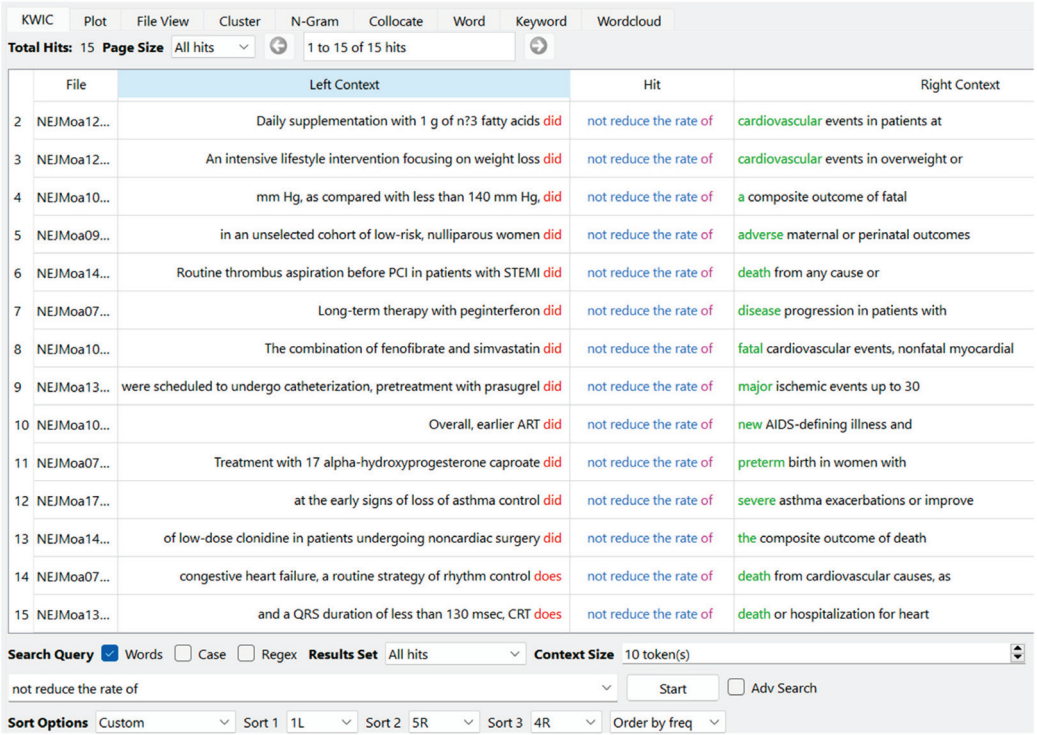

Figure 3. Example screen of AntConc showing "not reduce the rate of" as the node word with 13 instances of "did" and 2 instances of "does" at L1 or one word to the left (order by frequency of L1).

5. Results

5.1. Lexical Bundles in Move Texts

5.1.1. Frequencies

We found a total of 1286 five-word lexical bundles in medical research abstracts: 71 in Move 1, 305 in Move 2, 848 in Move 3, and 62 in Move 4. The most frequently occurring bundles are shown in Tables 4–7, and all 1286 bundles are shown in Table A1 in Appendix A. The most frequent five-word bundle "the efficacy and safety of" appeared 98 times in Move 1 with a range of 96; "the primary end point was" in Move 2 occurred 708 times in 705 texts. The bundle "confidence interval ci to p" occurred 681 times in 681 texts in Move 3. The original sequence was "confidence interval [CI], n# to n#; P<n#" as shown in an example sentence below, after the exclusion of punctuation and digits (Durrant 2017, p. 170; Viana and O'Boyle 2022, p. 119). In Move 4, the most frequently occurring bundle "associated with an increased risk" appeared 59 times in 56 texts. These instances indicate that many high-frequency sequences were used only once per abstract. The raw frequency and range values suggest a consistent use of bundles across the corpus texts.

Table 4. Most frequent five-word lexical bundles in Move 1.

Rank	Five-Word Lexical Bundle	Freq	Range
1	the efficacy and safety of	98	96
2	non small cell lung cancer	47	47
3	the safety and efficacy of	45	45
4	low density lipoprotein ldl cholesterol	42	42
4	with an increased risk of	42	41
6	human immunodeficiency virus type hiv	40	40
6	associated with an increased risk of	40	39
7	in patients with type diabetes	37	32
8	chronic obstructive pulmonary disease copd	31	31
8	coronary artery bypass grafting cabg	31	31
8	it is not known whether	31	31

Table 5. Most frequent five-word lexical bundles in Move 2.

Rank	Five-Word Lexical Bundle	Freq	Range
1	the primary end point was	708	705
2	at a dose of mg	325	252
3	we randomly assigned patients with	294	294
4	were randomly assigned to receive	285	278
5	primary end point was the	278	277
6	per kilogram of body weight	211	211
7	the primary outcome was the	190	190
8	in a ratio to receive	183	178
9	randomly assigned in a ratio	151	147
10	patients were randomly assigned to	149	148

Table 6. Most frequent five-word lexical bundles in Move 3.

Rank	Five-Word Lexical Bundle	Freq	Range
1	confidence interval ci to p	681	681
2	hazard ratio ci to p	421	272
3	percent confidence interval to p	310	155
4	of the patients in the	291	192
5	hazard ratio confidence interval ci	230	230
6	in of patients in the	225	149
7	than in the placebo group	221	173
8	similar in the two groups	208	193
9	in the placebo group p	185	138
10	a total of patients were	180	180
10	and in the placebo group	180	138

Table 7. Most frequent five-word lexical bundles in Move 4.

Rank	Five-Word Lexical Bundle	Freq	Range
1	associated with an increased risk	59	56
2	with an increased risk of	58	55
3	did not result in a	38	38
4	at a dose of mg	31	27
5	was associated with an increased	30	30
6	of death from any cause	27	24
6	there was no significant difference	27	27
8	than among those who received	26	26
9	out of hospital cardiac arrest	23	22
9	was associated with a higher	23	23

1. In total, 457 patients (22.8%) in the surgery group and 539 patients (26.4%) in the control group died (hazard ratio, 0.77; 95% confidence interval [CI], 0.68 to 0.87; $p < 0.001$).

While the top bundle in Move 1 did not exceed a raw frequency of 100 (Table 4), the bundle "the primary end point was" in Move 2 was notably prevalent and was found 708 times in 705 texts (Table 5). The prevalence of this bundle indicates common practices in the disciplinary research where the sequence is used to define main variable or parameter to be measured. The combination reflects a highly technical aspect, setting the stage for the methodological framework although the individual lexical items—"the", "primary", "end", "point", and "was"—have been found to be accessible from a vocabulary perspective (Asano and Fujieda 2024). For instance, the following example illustrates the essential role of such lexical bundles in presenting the main parameters of a study in a succinct manner:

2. The primary end point was histologic improvement in the 10-mg group as compared with the placebo group.

The bundle "confidence interval ci to p" also occurred as high as 681 times in 681 texts in Move 3, as represented by the fragment "95% confidence interval [CI], 0.68 to 0.87; $p < 0.001$" in the abovementioned example sentence. This indicates a noticeably consistent appearance in statistical contexts. The abbreviation "CI" indicates multiple occurrences of the term "confidence interval" within the text. The consistent appearance of the five-word bundle "confidence interval ci to p" across various texts suggests its routine use in statistical analyses within medical research. This bundle helps in presenting critical statistical information in a uniform manner, which facilitates clear and comparable interpretations of study results. For instance, the text uses this bundle to report the statistical measures of effect and significance, such as the confidence interval and a *p*-value. The text illustrates the bundle's role in showing quantitative aspects of the study findings. This uniform application of statistical terminology ensures that the findings are communicated in a precise and standardized format that can be easily understood and evaluated by others in the scientific community. In Move 4, high-frequency bundles contained word items related to discussing changes or statistical results, such as "associated with an increased risk" and "there was no significant difference", as described later in Section 5.1.3. Structure and Functions of Lexical Bundles.

Figure 4 shows the frequency of the leading bundles in each move. This indicates that the individual bundles have unique roles in the texts, as shown by Nam et al. (2016). For example, Panel A in Figure 4 shows that "the efficacy and safety of" occurred most often in Move 1. Example concordance lines indicated that the bundle took the technical part of the sentence for realizing study objectives.

3. We assessed the efficacy and safety of a paclitaxel-coated balloon in this setting.

Panel B in Figure 4 indicates that the instances of "the primary end point was" were dominant in Move 2. The bundle "confidence interval ci to p" only occurred in Move 3 (Panel C), indicating its essential function for presenting statistical findings. The sequence "associated with an increased risk" was rather exceptional as it was common in Moves 3 and 4 and also occurred in Move 1. However, the instances in these moves were below 60; therefore, the bundle may have been used only when the bundle fitted appropriately in realizing the authors' intentions. The findings suggest that these combinations may have pedagogical value. Becoming aware of where these bundles typically appear and their function within the abstracts could greatly enhance learners' familiarity with disciplinary research article abstracts.

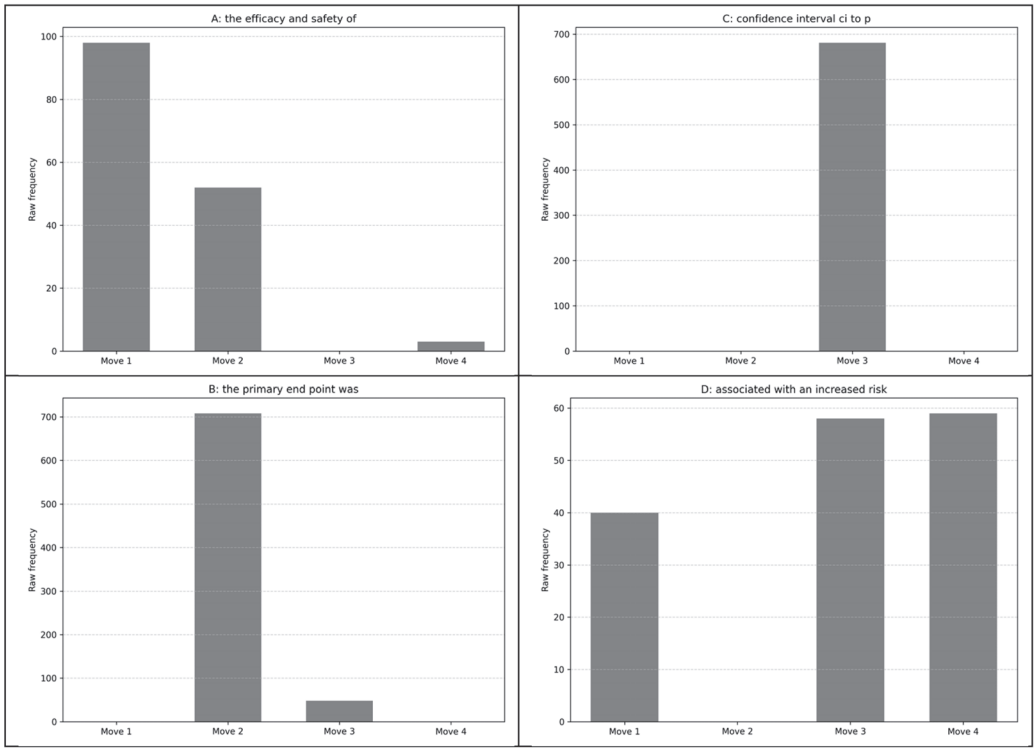

Figure 4. Frequency of move-specific most frequent five-word bundles. (**A**): the efficacy and safety of; (**B**): the primary end point was; (**C**): confidence interval ci to p; (**D**): associated with an increased risk.

5.1.2. The NSGL Coverage of the Bundles

The coverage of the first thousand NGSL varied across the moves (Figure 5). The word list covered up to 83.9% of the word items in bundles occurred in Move 4, followed by 76.4% in Move 3, 66.7% in Move 2, and 54.6% in Move 1. These variations are explicitly depicted in Figure 6. This figure shows the percentages of word items in each five-word bundle covered by the word list. For instance, a five-word bundle was classified as 100% when all five individual word items were covered by the word list. It was marked as 80% when four out of the five word items were covered. When none of the word items were covered by the word list, the bundle was marked as 0%.

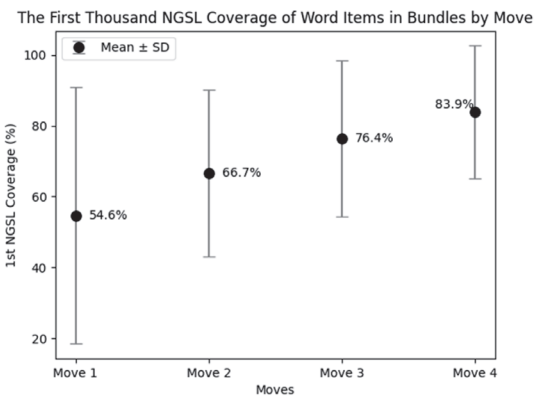

Figure 5. The first thousand NGSL coverage of word items in lexical bundles by move.

In Move 4, about a half of the bundles consisted of words in the first NGSL only, such as "with an increased risk of", "did not result in a", and "was not associated with a". The coverage of the word list was 40% or greater in all the bundles in Move 4. These bundles were often used for discussing the changes or statistical results.

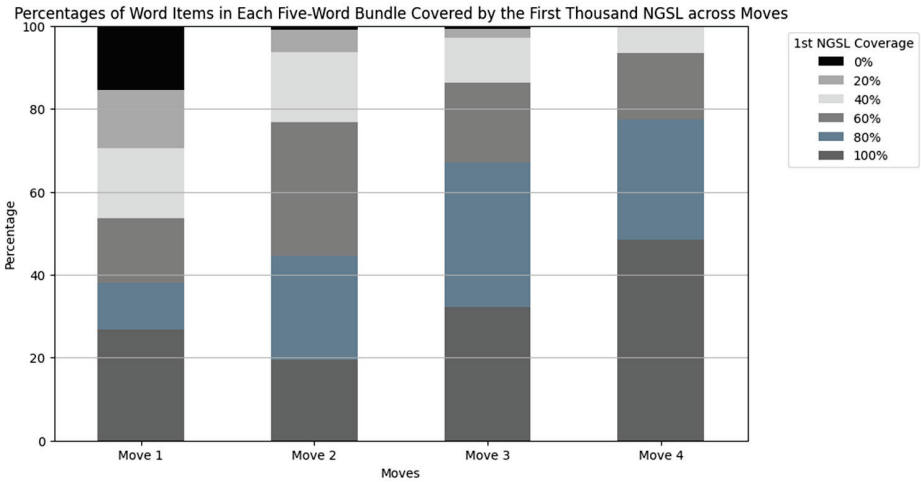

Figure 6. Percentages of word items in each five-word bundle covered by the first thousand NGSL across moves. From the bottom, bars in dark gray represent 100%, those in blue gray represent 80%, those in dim gray represent 60%, those in silver represent 40%, those in light gray represent 20%, and those in black represent 0%.

On the other hand, 15.5% of the bundles in Move 1 contained no word items covered by the word list, such as "hepatitis c virus hcv infection" (Table 8). Additionally, 14.1% of the sequences had only one word item covered by the word list, such as the fourth most frequently occurring bundle "low density lipoprotein ldl cholesterol" (Table 4). These combinations were used to introduce the research area as described in the following Section 5.1.3.

Table 8. Concordance lines with the node word of "hepatitis C virus (HCV) infection" in Move 1.

chronic	hepatitis C virus (HCV) infection	is a cause of major complications
chronic	hepatitis C virus (HCV) infection	is more prevalent among
therapy for chronic	hepatitis C virus (HCV) infection	is effective in less than 50% of
in patients with chronic	hepatitis C virus (HCV) infection.	We compared the efficacy and
for patients with chronic	hepatitis C virus (HCV) infection.	We evaluated daclatasvir (an
ribavirin for chronic	hepatitis C virus (HCV) infection.	However, these regimens have
Patients with chronic	hepatitis C virus (HCV) infection	who have not had a response to
response to therapy for	hepatitis C virus (HCV) infection,	and limited data suggest that
facilitate treatment for	hepatitis C virus (HCV) infection	in patients with thrombocytopenia
standard treatment for	hepatitis C virus (HCV) infection	is interferon, which is administered

Move 3 also had many bundles comprised of words covered by the word list, but they were different from the bundles in Move 4 in that most of the bundles in Move 3 were used to accurately report results with numerals, such as "percent confidence interval to p", "of the patients in the", and "than in the placebo group" (Table 6). Although each word item was rather accessible, the bundles appeared to convey information specific to the research (Table 9).

Table 9. Concordance lines with the node word of "of the patients in the" in Move 3.

centers: 6.7 percent	of the patients in the	angioplasty group reached the primary
was reached in 8.5 percent	of the patients in the	angioplasty group, as compared with
performed in 12.0 percent	of the patients in the	paclitaxel-stent group and 6.4
angiography in 16.5 percent	of the patients in the	paclitaxel-stent group and 6.9
group and 18 percent	of the patients in the	placebo group reported
severity index in 4 percent	of the patients in the	placebo group, as compared with
group and 6.4 percent	of the patients in the	sirolimus-stent group (p = 0.13).
Only 42 percent	of the patients in the	intraperitoneal-therapy group
Overall, 12 percent	of the patients in the	lamivudine group and 18 percent
index in 25 percent	of the patients in the	low-dose group, 44 percent of

Similar tendencies were seen in Move 2, where many frequent bundles, such as "the primary end point was" and "we randomly assigned patients with", were embedded in the context of research as shown in the next section. There were 31 instances of the first-person pronoun, and they were generally used to build research-oriented bundles. Here again, while the individual word items seemed accessible, the combined use of the words tended to reflect the disciplinary conventions.

5.1.3. Structure and Functions of Lexical Bundles

The texts showed that main structures of bundles were noun phrases, similar to the findings in academic written discourse such as those by Biber et al. (1999), Hyland (2008a), and Golparvar and Barabadi (2020). Noun phrases most often occurred in Move 1 (Table 10). Many were used for presenting disciplinary content words like "low-density lipoprotein (LDL) cholesterol" and "chronic obstructive pulmonary disease (COPD)" for setting the scene:

4. Non–small-cell lung cancer with sensitive mutations of the epidermal growth factor receptor (EGFR) is highly responsive to EGFR tyrosine kinase inhibitors such as gefitinib, but little is known about how its efficacy and safety profile compares with that of standard chemotherapy.

Table 10. Principal structures of bundles across moves (%).

Structure	Move 1	Move 2	Move 3	Move 4
NP-based	**57.7**	**39.0**	**48.1**	**41.9**
Noun phrase	15.5	5.9	2.4	3.2
Noun phrase + of	9.9	7.9	13.3	16.1
Noun phrase + other	32.4	25.2	32.4	22.4
PP-based	**14.1**	**20.7**	**25.6**	**21.0**
Prepositional phrase + of	2.8	7.2	5.7	12.9
Other prepositional phrase	11.3	13.4	19.9	8.1
VP-based	**26.8**	**39.3**	**24.9**	**30.6**
Passive verb	4.2	3.9	2.7	12.9
Pronoun/noun + verb	11.3	11.1	5.4	0.0
Pronoun/noun + be	2.8	13.8	7.7	1.6
Verb + noun/prep. phrase	5.6	6.2	5.2	16.1
Copula be + noun/adjective phrase	1.4	3.0	3.9	0.0
to-clause fragment	1.4	1.3	0.0	0.0
Others	**0.0**	**1.0**	**1.4**	**6.5**
Anticipatory it structure	**1.4**	**0.0**	**0.0**	**0.0**
Totals	100.0	100.0	100.0	100.0

On the other hand, *of*-phrase fragments such as "a significantly lower rate of" appeared frequently in expressing the conclusions (Move 4). Most sequences were part of descriptions used to indicate comparisons:

5. In children with chronic hepatitis B, 52 weeks of treatment with lamivudine was associated with a significantly higher rate of virologic response than was placebo.

The results (Move 3) made by far the most use of bundles beginning with a prepositional phrase. Such items included "of the patients in the", "in the medical therapy group", and "as compared with the placebo". These were most likely used to describe classifications and comparisons:

6. At 3 years, the criterion for the primary end point was met by 5% of the patients in the medical-therapy group, as compared with 38% of those in the gastric-bypass group ($p < 0.001$) and 24% of those in the sleeve-gastrectomy group ($p = 0.01$).

Verb phrases were most used for describing methods (Move 2). These included passive bundles preceded by "patients", such as "patients were randomly assigned to", and constructions with the presence of the first-person pronoun, such as "we evaluated the efficacy of". The following is an example with the first-person pronoun:

7. We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with COVID-19 and had evidence of lower respiratory tract infection.

There were very few usages of the anticipatory it structure among five-word bundles. The only sequence we found was "it is not known whether" in Move 1:

8. It is not known whether infants conceived with use of intracytoplasmic sperm injection or in vitro fertilization have a higher risk of birth defects than infants conceived naturally.

Many sequences were research-oriented bundles (Table 11). These bundles were used to introduce the research area in Move 1, such as "coronary artery bypass grafting cabg". The specific word combinations were used and defined to engage readers and direct their focus to the research areas early in the abstracts.

Table 11. Distribution of bundle functions by move (%).

Function	Move 1	Move 2	Move 3	Move 4
Research-oriented	87.3	98.7	91.3	62.9
Text-oriented	4.2	1.3	8.4	37.1
Participant-oriented	8.5	0.0	0.4	0.0
Totals	100.0	100.0	100.0	100.0

9. Some studies suggest that combination antiretroviral therapy in pregnant women with human immunodeficiency virus type 1 (HIV-1) infection increases the risk of premature birth and other adverse outcomes of pregnancy.

Almost all bundles were shown to describe study-related matters in Move 2. They were used to describe "location—indicating time/place" (Hyland 2008a, p. 13) such as "at the time of the", procedure such as "we randomly assigned to a", and quantification such as "a scale from to with". These bundles functioned to contextualize the research processes.

10. Each videotape was rated in various domains of technical skill on a scale of 1 to 5 (with higher scores indicating more advanced skill) by at least 10 peer surgeons who were unaware of the identity of the operating surgeon.

Although the results section (Move 3) was also predominantly composed of research-oriented bundles, it showed a few participant-oriented bundles like "were more likely to be". Such bundles were accompanied by specific numbers, or values, and results of statistical analyses. The sequences helped to support the summarized data to describe major findings.

11. According to univariate analysis, patients with *S. marcescens* bacteremia stayed in the surgical intensive care unit longer than controls (13.5 vs. 4.0 days, $p < 0.001$), were more likely to have received fentanyl in the surgical intensive care unit (odds ratio, 31; $p < 0.001$), and were more likely to have been exposed to two particular respiratory therapists (odds ratios, 13.1 and 5.1; $p < 0.001$ for both comparisons).

Move 4 showed outstanding differences in that it had various text-oriented bundles. They were related to changes seen in the studies in relation to certain interventions, such as "associated with an increase in" or indicating statistical results, such as "no significant difference in the". These bundles helped briefly discuss and conclude the study findings in abstract texts.

12. Knowledge of the fetal oxygen saturation is not associated with a reduction in the rate of cesarean delivery or with improvement in the condition of the newborn.

The instances of "participant-oriented bundles" (Hyland 2008a, p. 14) were fewer than expected, indicating a preference for diverse word combinations to construct participant-oriented discourse. Specifically, the bundle "may reduce the risk of" and "may increase the risk of", which incorporate the modal "may", were extracted only in Moves 1 and 4, with more frequent occurrences in Move 1. The bundles such as "more likely to have" and "more likely to be" were found in Moves 3 and 4. These bundles were consistently paired with quantifiers like "more" or "less", likely reflecting the authors' intention to contextualize and interpret their results within comparative frameworks.

Despite the simplicity of the individual lexical items such as prepositions, pronouns, and verbs like "know", their combination into bundles plays significant roles in structuring academically contextualized sequences. This synergy illustrates how domain-specific language is formulated from simple components to convey complex ideas, which is crucial for precise communication in scientific contexts.

6. Discussion

Our main purpose in this study was to examine the lexical bundles across moves in the research article abstracts with a focus on the vocabulary levels. We aimed to address the two research questions. Our findings on the most frequently occurring five-word bundles in each move revealed many sequences similar to the 185 bundles extracted by Qi and Pan (2020) but showed a substantial difference in numbers. We identified 1286 five-word lexical bundles in total; about two-thirds of them occurred in Move 3, one-quarter in Move 2, and the remaining bundles were in either Move 1 or Move 4.

Our findings on the structures and functions of bundles support studies by Biber et al. (1999) and Hyland (2008a), which showed that the main structures of bundles are noun phrases, with the majority of sequences being research-oriented bundles. We have extended these studies by examining the coverage of the first thousand NGSL for bundles in each move. We found fluctuations in the coverage of the first thousand NGSL among bundles in different moves.

Although the word list covered around 80% of the word items in the bundles in Move 3 and Move 4, differences were seen in the forms and functions of the bundles in these moves. In Move 3, noun-phrase bundles accounted for about 50%, and many sequences were used to summarize research findings, describing essential data and presenting statistical information. In Move 4, however, about one-third of the bundles were verb phrases, and text-oriented bundles comprised 40%. The bundles in Move 2 were made up of word items in which the coverage of the word list was slightly lower than in the Move 3 bundles and were predominantly research-oriented. The first-person pronoun used in the Move 2 bundles was mainly part of research-oriented sequences. These pronouns were used to describe the researchers' adherence to disciplinary requirements (Hyland 2008a). Although many five-word lexical bundles were built with basic word items, the combined sequences often realized the contexts of research. Becoming familiar with these bundles may require learners to raise their awareness of the contextual knowledge of formal and rhetorical features of the genre texts (Tardy 2009). These findings may have implications for teaching specific lexical bundles in educational settings especially for novice learners such as undergraduate medical students.

Our study has several limitations. One is that we identified lexical bundles primarily based on frequency information. Nation (2001) argued that items with higher frequency are used more often and thus have educational value. Biber et al. (2004) suggested that examination of frequently occurring bundles can reveal the real-world use of multi-word sequences. However, there are counterarguments. Flowerdew (2012), citing Widdowson (1991) and Cook (1998), pointed out "the danger of equating frequency with pedagogic relevance" (p. 191). This is pertinent because Simpson-Vlach and Ellis (2010) argued that "sequences such as 'on the other hand' and 'at the same time' are more psycholinguistically salient than sequences such as 'to do with the', or 'I think it was'" (p. 490). They created an

academic formula list by taking into consideration the psycholinguistic salience in addition to frequency count.

A notable finding from Graetz (1982) is that there is a high correlation between the disciplines in which abstracts appear in journals and lecturers' reliance on abstracts, along with their belief that students should be taught how to read abstracts (p. 26). This perspective is supported by Nation (2022), who emphasized that the characteristics and usage of language items should substantially influence teaching and learning strategies (p. 435). However, our study primarily focused on analyzing the features of lexical bundles in abstract texts, without exploring into classroom applications.

In one of our prior studies, we used a seven-year segment of our corpus to develop the Medical English Education Support System (MEESUS), which was implemented in language courses at a university (Asano et al. 2022a). The complexity of contextual meanings of lexical items in disciplinary discourse has been pointed out, such as the term "clinical" in "clinical trials" and "a clinical decision" (Coxhead 2016, p. 179). To tackle this difficulty, around 100 first-year "genre students" (Hyland 2008a, p. 20), with an average TOEFL ITP score of 475.04, engaged with the MEESUS in our previous study (Asano et al. 2022a). They used its concordance tool to explore language items and reported their insights on terms like "mean" for an average and "case" for a subject, gaining awareness of their their usage in the academic contexts.

In another study, fourth-year medical students, averaging a TOEFL ITP score of 455.9, used disciplinary "guidelines" (Millar et al. 2019, p. 150) to examine and summarize research abstracts (Asano et al. 2022b). Despite different thoughts about "learner-directed corpus projects" (Ballance and Coxhead 2022, p. 412), such activities have been shown to enhance linguistic "awareness" and "tolerance", preparing learners for real-world applications (Cook 2010, pp. 117–18). Simpson-Vlach and Ellis (2010) suggested that "organization of constructions according to academic needs and purposes is essential" in transforming language resources into effective tools for pedagogy (p. 510).

In the future, it will be necessary to expand upon these prior works by conducting classroom activities. Such activities may include observing lexical bundles and raising learners' awareness of the form and function of the sequences in context.

Many studies have shown the linguistic features of lexical items in disciplinary texts (Biber et al. 1999; Cortes 2004, 2013; Hyland 2008a, 2008b; Mizumoto et al. 2017; Omidian et al. 2018; Qi and Pan 2020; Li et al. 2020). These studies have aimed to address the challenges faced by "novice and seasoned scientists" (Kanoksilapatham 2005, p. 288). This includes the notion that becoming proficient in a language requires an awareness of the preferred word sequences used by experts (Hyland 2008b, p. 44). Our study found similar results regarding five-word lexical bundles in the moves of research article abstracts (Biber et al. 1999; Hyland 2008a). Our examination revealed that although the bundles identified in our texts consisted of relatively accessible individual vocabulary items, their combined sequences reflected the context of research and adherence to the required academic conventions. Several difficulties in learning multiword items have been pointed out (Boers 2020). Becoming familiar with these patterns could assist students in "learning productive chunks of language" (Reppen and Olson 2020, p. 177). In helping learners succeed in their learning contexts (Tribble 2017), the findings in this study suggest that raising awareness of specific lexical bundles can be beneficial for aiding students in joining their disciplinary communities.

Author Contributions: Conceptualization, M.A. and M.F.; text processing and corpus compilation K.H.; analyses, M.A., K.H. and M.F.; writing—original draft preparation, M.A., K.H. and M.F.; funding acquisition, M.A. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Grants-in-Aid for Scientific Research awarded by the Japan Society for the Promotion of Science (JSPS) grant numbers 23K02800 and 18K02966. The APC was funded by the Author Voucher discount code (0736e63e42811a20).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data from our corpus are currently not accessible to the public. However, data used in this study can be provided upon request to the corresponding author.

Acknowledgments: We would like to express our gratitude to the two anonymous reviewers for their insightful comments. We also thank the academic editors of *Languages* for their generous assistance throughout the review process. The authors are especially grateful to Junichiro Taki, a fourth-year medical student at Osaka Medical and Pharmaceutical University, for his invaluable assistance in data processing and corpus compilation. Any remaining errors are our responsibility.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

All lexical bundles identified are listed by move in the order of frequency (Table A1). The abbreviations are as follows: Freq stands for frequency, 1st N for the coverage of the first NGSL, NP for noun phrase, VP for verb phrase, PP or prepositional phrase, Ant it for anticipatory it structure, Res for research-oriented, Text for text-oriented, and Participant for participant-oriented bundles.

Table A1. All lexical bundles identified by move, frequency, first NGSL coverage, structure, and function.

Move	Bundles	Freq	Range	1st N	Structure	Function
1	the efficacy and safety of	98	96	60	NP	Res
1	non small cell lung cancer	47	47	40	NP	Res
1	the safety and efficacy of	45	45	60	NP	Res
1	low density lipoprotein ldl cholesterol	42	42	20	NP	Res
1	with an increased risk of	42	41	100	PP	Res
1	human immunodeficiency virus type hiv	40	40	100	NP	Res
1	associated with an increased risk	40	39	40	VP	Res
1	in patients with type diabetes	37	32	80	PP	Res
1	chronic obstructive pulmonary disease copd	31	31	20	NP	Res
1	coronary artery bypass grafting cabg	31	31	0	NP	Res
1	it is not known whether	31	31	100	Ant it	Text
1	small cell lung cancer nsclc	30	30	40	NP	Res
1	human immunodeficiency virus hiv infection	28	28	20	NP	Res
1	we tested the hypothesis that	27	27	80	VP	Res
1	little is known about the	26	26	100	VP	Participant
1	out of hospital cardiac arrest	23	21	60	NP	Res
1	are at high risk for	22	22	100	VP	Res
1	human epidermal growth factor receptor	22	22	60	NP	Res
1	we conducted a study to	22	22	80	VP	Res
1	years of age or older	21	20	100	NP	Res
1	we sought to determine whether	20	20	100	VP	Res
1	hepatitis c virus hcv infection	19	19	0	NP	Res
1	of human immunodeficiency virus hiv	19	19	40	PP	Res
1	influenza a h n virus	18	18	20	NP	Res
1	high density lipoprotein hdl cholesterol	17	17	20	NP	Res
1	is a major cause of	17	17	100	VP	Participant
1	the human immunodeficiency virus hiv	17	17	40	NP	Res
1	acute respiratory distress syndrome ards	16	16	0	NP	Res
1	hepatitis c virus hcv genotype	16	16	0	NP	Res
1	in patients with atrial fibrillation	16	15	60	PP	Res
1	to reduce the risk of	16	16	100	PP	Res
1	we conducted a randomized trial	16	16	40	VP	Res
1	are at increased risk for	15	15	100	VP	Res
1	has been shown to reduce	15	15	100	VP	Text
1	proprotein convertase subtilisin kexin type	15	15	20	NP	Res
1	transcatheter aortic valve replacement tavr	15	15	0	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
1	we evaluated the effect of	15	15	80	VP	Res
1	with hepatitis c virus hcv	15	15	20	PP	Res
1	epidermal growth factor receptor egfr	14	14	40	NP	Res
1	has been shown to be	14	13	100	VP	Text
1	may reduce the risk of	14	14	100	VP	Participant
1	we evaluated the efficacy of	14	14	60	VP	Res
1	we evaluated the safety and	14	14	60	VP	Res
1	with human immunodeficiency virus hiv	14	14	40	PP	Res
1	diffuse large b cell lymphoma	13	9	40	NP	Res
1	st segment elevation myocardial infarction	13	13	0	NP	Res
1	the treatment of patients with	13	13	100	NP	Res
1	allogeneic hematopoietic stem cell transplantation	12	12	20	NP	Res
1	angiotensin converting enzyme ace inhibitors	12	12	0	NP	Res
1	in patients with heart failure	12	12	80	PP	Res
1	primary percutaneous coronary intervention pci	12	12	0	NP	Res
1	the most common cause of	12	12	100	NP	Participant
1	we assessed the efficacy and	12	12	60	VP	Res
1	cardiovascular events in patients with	11	10	80	NP	Res
1	continuous positive airway pressure cpap	11	11	40	NP	Res
1	cystic fibrosis transmembrane conductance regulator	11	11	0	NP	Res
1	has not been well studied	11	11	100	VP	Participant
1	methicillin resistant staphylococcus aureus mrsa	11	11	0	NP	Res
1	patients with chronic kidney disease	11	9	60	NP	Res
1	patients with relapsed or refractory	11	10	60	NP	Res
1	patients with type diabetes mellitus	11	11	60	NP	Res
1	severe acute respiratory syndrome coronavirus	11	11	0	NP	Res
1	the long term effects of	11	11	100	NP	Res
1	the risk of cardiovascular events	11	11	80	NP	Res
1	we examined the effect of	11	11	80	VP	Res
1	with acute myeloid leukemia aml	11	11	20	PP	Res
1	data are lacking on the	10	10	100	VP	Participant
1	graft versus host disease gvhd	10	10	20	NP	Res
1	mutations in the gene encoding	10	10	40	NP	Res
1	of death from any cause	10	10	100	PP	Res
1	patients with severe aortic stenosis	10	10	40	NP	Res
2	the primary end point was	708	705	80	VP	Res
2	at a dose of mg	325	252	60	PP	Res
2	we randomly assigned patients with	294	294	60	VP	Res
2	were randomly assigned to receive	285	278	60	VP	Res
2	primary end point was the	278	277	80	VP	Res
2	per kilogram of body weight	211	211	80	PP	Res
2	the primary outcome was the	190	190	60	VP	Res
2	in a ratio to receive	183	178	80	PP	Res
2	randomly assigned in a ratio	151	147	40	VP	Res
2	patients were randomly assigned to	149	148	60	VP	Res
2	randomized double blind placebo controlled	134	134	20	NP	Res
2	assigned in a ratio to	127	123	60	VP	Res
2	to with higher scores indicating	125	105	100	PP	Res
2	years of age or older	125	121	100	NP	Res
2	were randomly assigned in a	124	120	60	VP	Res
2	double blind placebo controlled trial	120	120	20	NP	Res
2	was a composite of death	93	88	80	VP	Res
2	the primary outcome was a	90	90	60	VP	Res
2	outcome was a composite of	79	77	60	VP	Res
2	we randomly assigned patients to	79	79	60	VP	Res
2	per square meter of body	75	75	60	PP	Res
2	we conducted a randomized double	72	72	40	VP	Res
2	the primary efficacy end point	71	70	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
2	we randomly assigned patients who	70	70	60	VP	Res
2	secondary end points included the	53	53	80	VP	Res
2	with the use of a	53	52	100	PP	Res
2	with the use of the	53	50	100	PP	Res
2	the efficacy and safety of	52	51	60	NP	Res
2	scores range from to with	51	42	100	NP	Res
2	of death from any cause	50	47	100	PP	Res
2	reverse transcriptase polymerase chain reaction	50	50	0	NP	Res
2	with higher scores indicating more	45	43	100	PP	Res
2	a total of patients with	44	44	100	NP	Res
2	between the ages of and	44	43	100	PP	Res
2	to years of age with	44	43	100	PP	Res
2	the primary end points were	43	43	80	VP	Res
2	we conducted a double blind	43	43	40	VP	Res
2	in this randomized double blind	41	41	40	PP	Res
2	mg per deciliter mmol per	40	29	40	NP	Res
2	the primary outcome measure was	39	39	60	VP	Res
2	to years of age who	39	39	100	PP	Res
2	we conducted a multicenter randomized	38	38	40	VP	Res
2	double blind placebo controlled phase	37	37	20	NP	Res
2	of death from cardiovascular causes	37	33	80	PP	Res
2	the primary efficacy outcome was	37	37	40	VP	Res
2	a double blind placebo controlled	34	33	40	NP	Res
2	change from baseline in the	34	31	80	NP	Res
2	the primary outcome was death	34	34	60	VP	Res
2	was death from any cause	34	34	100	VP	Res
2	were to years of age	34	33	100	VP	Res
2	were randomly assigned to undergo	33	33	40	VP	Res
2	ml per minute per m	31	23	80	NP	Res
2	this double blind placebo controlled	31	31	40	NP	Res
2	at a dose of μg	30	24	60	PP	Res
2	forced expiratory volume in second	30	30	60	NP	Res
2	in a randomized double blind	30	30	40	PP	Res
2	on the basis of the	29	29	100	PP	Text
2	outcome was the rate of	29	27	80	VP	Res
2	were years of age or	29	29	100	VP	Res
2	assessed with the use of	28	27	80	VP	Res
2	end point was the percentage	28	28	80	VP	Res
2	the presence or absence of	28	28	60	NP	Text
2	we randomly assigned patients in	28	28	60	VP	Res
2	assigned to receive mg of	27	26	60	VP	Res
2	or death from cardiovascular causes	26	24	80	NP	Res
2	the primary safety end point	26	24	60	NP	Res
2	trial we assigned patients with	26	26	60	VP	Res
2	double blind randomized placebo controlled	25	25	20	NP	Res
2	mg per day or placebo	25	24	60	NP	Res
2	nonfatal myocardial infarction nonfatal stroke	25	24	0	NP	Res
2	double blind trial we randomly	24	24	20	NP	Res
2	placebo controlled trial we randomly	24	24	40	NP	Res
2	on the modified rankin scale	23	20	40	PP	Res
2	to receive either mg of	23	23	80	VP	Res
2	a total of patients were	22	22	100	NP	Res
2	end point was death from	22	22	100	VP	Res
2	placebo controlled phase trial we	22	22	40	NP	Res
2	the change from baseline to	22	21	80	NP	Res
2	the two primary end points	22	22	80	NP	Res
2	we conducted a randomized trial	22	22	40	VP	Res
2	weight in kilograms divided by	22	22	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
2	divided by the square of	21	21	60	VP	Res
2	patients with moderate to severe	21	21	60	NP	Res
2	performed with the use of	21	21	100	VP	Res
2	randomized double blind trial we	21	21	20	NP	Res
2	secondary end points included overall	21	21	60	NP	Res
2	the patients were randomly assigned	21	21	60	VP	Res
2	was a sustained virologic response	21	21	60	VP	Res
2	weeks after the end of	21	21	100	NP	Res
2	assigned patients with type diabetes	20	20	60	VP	Res
2	at a dose of or	20	17	80	PP	Res
2	composite of death myocardial infarction	20	19	40	NP	Res
2	end point was a sustained	20	20	80	VP	Res
2	health related quality of life	20	20	100	NP	Res
2	in the score on the	20	18	100	PP	Res
2	nonfatal myocardial infarction or nonfatal	20	20	20	NP	Res
2	out of hospital cardiac arrest	20	18	60	NP	Res
2	we randomly assigned adults with	20	20	60	VP	Res
2	children to months of age	19	19	100	NP	Res
2	end points were overall survival	19	19	60	VP	Res
2	free survival and overall survival	19	19	40	NP	Res
2	in the intention to treat	19	18	80	PP	Res
2	on a scale of to	19	17	80	PP	Res
2	or placebo in addition to	19	19	80	NP	Res
2	outcome was the composite of	19	18	60	VP	Res
2	to mg per deciliter to	19	16	60	PP	Res
2	a time to event analysis	18	15	100	NP	Res
2	at a dose of to	18	16	80	PP	Res
2	double blind phase trial we	18	18	20	NP	Res
2	ejection fraction of or less	18	18	60	NP	Res
2	randomly assigned to one of	18	18	60	VP	Res
2	the primary composite end point	18	18	60	NP	Res
2	was the composite of death	18	17	80	VP	Res
2	after the end of treatment	17	17	100	PP	Res
2	at the end of the	17	16	100	PP	Res
2	end point was the composite	17	15	80	VP	Res
2	hours after the onset of	17	17	80	NP	Res
2	in this phase trial we	17	17	60	PP	Res
2	investigator assessed progression free survival	17	17	20	NP	Res
2	mg per kilogram every weeks	17	10	60	NP	Res
2	non small cell lung cancer	17	16	40	NP	Res
2	patients were assigned to receive	17	15	80	VP	Res
2	progression free survival and overall	17	17	40	NP	Res
2	we conducted a retrospective cohort	17	17	40	VP	Res
2	were randomly assigned to a	17	17	60	VP	Res
2	a left ventricular ejection fraction	16	16	20	NP	Res
2	a scale from to with	16	15	80	NP	Res
2	key secondary end point was	16	16	80	VP	Res
2	mg twice daily or placebo	16	15	40	NP	Res
2	or death from any cause	16	16	100	NP	Res
2	randomization was stratified according to	16	16	60	VP	Res
2	randomized placebo controlled double blind	16	16	20	NP	Res
2	secondary end points were the	16	16	80	VP	Res
2	the between group difference in	16	14	100	NP	Text
2	the coprimary end points were	16	16	80	VP	Res
2	the median follow up was	16	16	80	VP	Res
2	the primary end point of	16	16	80	NP	Res
2	the safety and efficacy of	16	16	60	NP	Res
2	the time to the first	16	14	100	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
2	to with lower scores indicating	16	14	100	PP	Res
2	we conducted an open label	16	16	60	VP	Res
2	we enrolled patients who had	16	16	80	VP	Res
2	who had not previously received	16	16	80	VP	Res
2	a composite of death myocardial	15	15	60	NP	Res
2	according to the intention to	15	15	80	PP	Res
2	analyzed with the use of	15	13	80	VP	Res
2	during the period from through	15	14	100	PP	Res
2	end points were progression free	15	15	80	VP	Res
2	every weeks for up to	15	14	100	NP	Res
2	in this multicenter double blind	15	15	40	PP	Res
2	of to mg per deciliter	15	13	60	PP	Res
2	open label phase trial we	15	15	40	NP	Res
2	the composite of death from	15	15	80	NP	Res
2	the percentage of patients with	15	13	80	NP	Res
2	the primary safety outcome was	15	15	40	VP	Res
2	a composite of cardiovascular death	14	12	60	NP	Res
2	after the end of therapy	14	14	80	PP	Res
2	as compared with placebo in	14	14	80	PP	Res
2	at the time of the	14	13	100	PP	Res
2	death myocardial infarction or stroke	14	14	40	NP	Res
2	free survival as assessed by	14	14	60	NP	Res
2	in a ratio to undergo	14	14	60	PP	Res
2	mg per kilogram per day	14	11	60	NP	Res
2	or hospitalization for heart failure	14	14	60	NP	Res
2	patients who had a response	14	13	100	NP	Res
2	to years of age and	14	14	100	PP	Res
2	was the percentage of patients	14	14	80	VP	Res
2	we conducted a randomized controlled	14	14	60	VP	Res
2	we randomly assigned women with	14	14	60	VP	Res
2	a composite of death or	13	13	80	PP	Res
2	and randomly assigned them to	13	13	60	VP	Res
2	bmi the weight in kilograms	13	13	60	NP	Res
2	boundary of the confidence interval	13	13	40	NP	Res
2	end point was disease free	13	13	100	VP	Res
2	end point was the first	13	9	100	VP	Res
2	in a double blind fashion	13	13	40	PP	Res
2	in this double blind phase	13	13	40	PP	Res
2	key secondary end points were	13	13	80	VP	Res
2	of body surface area and	13	13	80	PP	Res
2	patients with hcv genotype infection	13	10	40	NP	Res
2	patients with relapsed or refractory	13	13	60	NP	Res
2	placebo for weeks the primary	13	12	60	NP	Res
2	point was investigator assessed progression	13	13	40	VP	Res
2	primary end point was disease	13	13	80	VP	Res
2	primary end point was survival	13	13	60	VP	Res
2	psoriasis area and severity index	13	13	40	NP	Res
2	randomized trial we assigned patients	13	13	40	NP	Res
2	sustained virologic response at weeks	13	13	60	NP	Res
2	time to event analysis was	13	12	100	NP	Res
2	to years of age to	13	12	100	PP	Res
2	trial to evaluate the efficacy	13	13	40	NP	Res
2	we performed a randomized double	13	13	60	VP	Res
2	were followed for up to	13	13	100	VP	Res
2	who did not have a	13	13	100	VP	Res
2	years of age or younger	13	13	100	NP	Res
2	a dose of either mg	12	11	60	NP	Res
2	aspirin at a dose of	12	11	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
2	at a daily dose of	12	11	80	PP	Res
2	at a dose of iu	12	10	60	PP	Res
2	between and years of age	12	12	100	PP	Res
2	body mass index bmi the	12	12	40	NP	Res
2	determined with the use of	12	12	100	VP	Res
2	end point was the annualized	12	12	80	VP	Res
2	estimated with the use of	12	12	100	VP	Res
2	evaluated with the use of	12	12	80	VP	Res
2	for a median of years	12	12	80	PP	Res
2	in this double blind trial	12	12	40	PP	Res
2	measured with the use of	12	12	100	VP	Res
2	mg once daily or placebo	12	12	60	NP	Res
2	of patients who had a	12	12	100	PP	Res
2	on the basis of a	12	12	100	PP	Text
2	real time polymerase chain reaction	12	12	40	NP	Res
2	secondary end point was the	12	12	80	VP	Res
2	st segment elevation myocardial infarction	12	12	0	NP	Res
2	than years of age who	12	12	100	PP	Res
2	the percentage of patients who	12	11	80	PP	Res
2	the primary composite outcome was	12	12	40	VP	Res
2	the primary objective was to	12	12	60	VP	Res
2	the primary outcome was day	12	12	60	VP	Res
2	the primary outcome was survival	12	12	40	VP	Res
2	the primary outcomes were the	12	12	60	VP	Res
2	the secondary end points were	12	12	80	VP	Res
2	to years of age in	12	11	100	PP	Res
2	trial in which patients with	12	12	80	NP	Res
2	was the time to the	12	11	100	VP	Res
2	we evaluated the effect of	12	12	80	VP	Res
2	we evaluated the efficacy of	12	12	60	VP	Res
2	we performed a multicenter randomized	12	12	60	VP	Res
2	were at high risk for	12	12	100	VP	Res
2	were randomly assigned to the	12	12	60	VP	Res
2	with a by factorial design	12	12	80	PP	Res
2	a by factorial design we	11	11	80	NP	Res
2	a noninferiority margin of percentage	11	11	40	NP	Res
2	an estimated glomerular filtration rate	11	11	60	NP	Res
2	an intention to treat basis	11	11	80	NP	Res
2	assessed in a time to	11	8	80	VP	Res
2	assigned to one of three	11	11	80	VP	Res
2	controlled trial involving patients with	11	11	80	PP	Res
2	death from coronary heart disease	11	10	80	PP	Res
2	disease patients were randomly assigned	11	11	60	VP	Res
2	dose of mg every weeks	11	11	60	NP	Res
2	dose of to mg per	11	10	60	NP	Res
2	factorial design we randomly assigned	11	11	40	NP	Res
2	from no symptoms to death	11	11	80	PP	Res
2	in the intensive care unit	11	11	80	PP	Res
2	in this double blind randomized	11	11	40	PP	Res
2	of less than mm hg	11	7	60	PP	Res
2	of the confidence interval for	11	11	60	PP	Res
2	open label randomized controlled trial	11	11	40	NP	Res
2	patients at high risk for	11	11	100	NP	Res
2	patients were stratified according to	11	11	80	VP	Res
2	patients who had had a	11	11	100	NP	Res
2	patients with type diabetes who	11	11	80	NP	Res
2	peginterferon alfa a and ribavirin	11	7	40	NP	Res
2	per square meter on day	11	8	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
2	randomly assigned them to receive	11	11	60	VP	Res
2	safety outcome was major bleeding	11	11	40	VP	Res
2	single nucleotide polymorphisms snps in	11	11	40	NP	Res
2	the principal safety outcome was	11	11	40	VP	Res
2	the rate of death from	11	10	100	NP	Res
2	to evaluate the efficacy and	11	11	60	VP	Res
2	to one of three groups	11	11	100	PP	Res
2	was a secondary end point	11	11	80	VP	Res
2	was death from cardiovascular causes	11	11	80	VP	Res
2	we randomly assigned patients years	11	11	60	VP	Res
2	who had a response to	11	11	100	VP	Res
2	who received a diagnosis of	11	10	80	VP	Res
2	within hours after symptom onset	11	11	60	PP	Res
2	a glycated hemoglobin level of	10	8	60	NP	Res
2	a multicenter double blind randomized	10	10	20	NP	Res
2	a multicenter randomized open label	10	10	40	NP	Res
2	and the primary end point	10	10	80	NP	Res
2	area under the curve auc	10	10	60	NP	Res
2	bone mineral density at the	10	8	40	NP	Res
2	cardiovascular death myocardial infarction or	10	10	40	NP	Res
2	cisplatin mg per square meter	10	10	20	NP	Res
2	clinical trial we randomly assigned	10	10	20	NP	Res
2	controlled study we randomly assigned	10	10	60	NP	Res
2	coprimary end points were the	10	10	80	VP	Res
2	data and safety monitoring board	10	10	60	NP	Res
2	death from any cause myocardial	10	10	80	NP	Res
2	estimated glomerular filtration rate gfr	10	10	40	NP	Res
2	evaluation criteria in solid tumors	10	10	20	NP	Res
2	had been randomly assigned to	10	9	60	VP	Res
2	less than copies per milliliter	10	10	60	Oth	Res
2	less than per cubic millimeter	10	9	60	Oth	Res
2	median follow up was years	10	10	80	VP	Res
2	noninferiority trial we randomly assigned	10	10	20	NP	Res
2	of at least mg per	10	8	80	PP	Res
2	older than years of age	10	10	100	Oth	Res
2	patients in the placebo group	10	10	80	NP	Res
2	proportional hazards models were used	10	10	60	VP	Res
2	ranges from to with higher	10	10	100	VP	Res
2	tested for an association between	10	9	80	VP	Res
2	the primary outcome measures were	10	10	60	VP	Res
2	to evaluate the efficacy of	10	10	60	VP	Res
2	to receive subcutaneous injections of	10	10	60	VP	Res
2	trial we enrolled patients with	10	10	60	NP	Res
2	we conducted a randomized open	10	10	60	VP	Res
2	we conducted a randomized placebo	10	10	40	VP	Res
2	we randomly assigned participants with	10	10	40	VP	Res
2	weeks days to weeks days	10	9	100	NP	Res
2	were assigned in a ratio	10	10	60	VP	Res
2	were randomly assigned to either	10	10	60	VP	Res
2	were randomly assigned to treatment	10	10	60	VP	Res
3	confidence interval ci to p	681	681	20	NP	Res
3	hazard ratio ci to p	421	272	20	NP	Res
3	percent confidence interval to p	310	155	40	NP	Res
3	of the patients in the	291	192	100	PP	Res
3	hazard ratio confidence interval ci	230	230	0	NP	Res
3	in of patients in the	225	149	100	PP	Res
3	than in the placebo group	221	173	80	Oth	Res
3	similar in the two groups	208	193	100	Oth	Text

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	in the placebo group p	185	138	60	PP	Res
3	a total of patients were	180	180	100	VP	Res
3	and in the placebo group	180	138	80	PP	Res
3	did not differ significantly between	176	163	60	VP	Text
3	a median follow up of	168	165	80	NP	Res
3	in of the patients in	168	114	100	PP	Res
3	occurred in patients in the	160	110	100	VP	Res
3	there was no significant difference	152	142	100	VP	Text
3	the placebo group hazard ratio	149	107	40	NP	Res
3	was confidence interval ci to	148	148	40	VP	Res
3	occurred in of the patients	145	119	100	VP	Res
3	were similar in the two	145	143	100	VP	Text
3	in the placebo group hazard	143	103	60	PP	Res
3	patients in the placebo group	142	105	80	NP	Res
3	as compared with in the	139	118	100	PP	Text
3	group hazard ratio ci to	138	109	40	NP	Res
3	there were no significant differences	134	128	100	VP	Text
3	percent percent confidence interval to	131	58	60	NP	Res
3	percent confidence interval to percent	123	62	60	NP	Res
3	mg per deciliter mmol per	121	60	40	NP	Res
3	occurred in of patients in	117	102	100	VP	Res
3	hazard ratio percent confidence interval	115	63	20	NP	Res
3	relative risk percent confidence interval	115	55	40	NP	Res
3	a total of patients underwent	109	109	80	VP	Res
3	confidence interval ci to and	107	107	40	NP	Res
3	difference percentage points ci to	104	70	60	NP	Res
3	group hazard ratio confidence interval	97	97	20	NP	Res
3	in the intention to treat	96	84	80	PP	Res
3	odds ratio ci to p	96	61	20	NP	Res
3	percentage points confidence interval ci	96	96	20	NP	Res
3	was similar in the two	95	89	100	VP	Text
3	percent confidence interval to and	94	71	60	NP	Res
3	confidence interval ci to in	93	93	40	NP	Res
3	difference percentage points confidence interval	90	90	40	NP	Res
3	ci to and ci to	89	67	60	NP	Res
3	relative risk ci to p	88	51	40	NP	Res
3	was percent confidence interval to	85	66	60	VP	Res
3	interval ci to p and	84	84	40	NP	Res
3	median progression free survival was	83	78	40	VP	Res
3	group as compared with the	79	74	100	NP	Text
3	vs hazard ratio ci to	79	44	20	PP	Res
3	than in the control group	78	60	100	Oth	Res
3	group and in in the	76	63	100	NP	Text
3	group as compared with of	74	65	100	NP	Text
3	group confidence interval ci to	74	74	40	NP	Res
3	group and patients in the	71	53	100	NP	Res
3	ratio ci to p and	71	59	40	NP	Res
3	the most common adverse events	71	71	80	NP	Res
3	ratio confidence interval to p	69	69	20	NP	Res
3	those in the placebo group	69	54	80	NP	Res
3	and of patients in the	68	60	100	PP	Res
3	risk confidence interval ci to	68	68	40	NP	Res
3	in the control group p	66	48	80	PP	Res
3	of death from any cause	66	61	100	PP	Res
3	years of age or older	66	54	100	NP	Res
3	a total of of the[3]	65	63	100	NP	Text
3	after a median follow up	65	64	80	PP	Res
3	in the placebo group and	64	50	80	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	the median progression free survival	64	60	40	NP	Res
3	there were no significant between	64	63	100	VP	Text
3	were randomly assigned to receive	63	57	60	VP	Res
3	group and percent in the	62	49	100	NP	Res
3	in the group that received	62	34	100	PP	Res
3	hazard ratio for disease progression	61	61	40	NP	Res
3	of the primary end point	61	56	80	PP	Res
3	patients were randomly assigned to	61	60	60	VP	Res
3	patients in the control group	60	40	100	NP	Res
3	respectively hazard ratio ci to	60	46	20	Oth	Res
3	the intention to treat analysis	60	58	80	NP	Res
3	and ci to in the	59	51	80	NP	Res
3	associated with an increased risk	58	54	100	VP	Res
3	death confidence interval ci to	58	58	40	NP	Res
3	in the medical therapy group	58	26	60	PP	Res
3	adverse events of grade or	57	54	60	NP	Res
3	as compared with of those	57	47	100	PP	Text
3	ci to p for noninferiority	57	51	40	NP	Res
3	group and the placebo group	56	45	80	NP	Res
3	group hazard ratio for death	56	53	60	NP	Res
3	to and ci to respectively	56	47	60	PP	Res
3	at a median follow up	55	53	80	PP	Res
3	of patients in the placebo	55	50	80	PP	Res
3	percentage points ci to p	54	41	40	NP	Res
3	adverse events were similar in	53	53	80	VP	Res
3	between group difference in the	53	50	100	NP	Res
3	for disease progression or death	53	52	80	PP	Res
3	of the patients who received	53	36	100	PP	Res
3	the median overall survival was	53	48	40	VP	Res
3	hazard ratio confidence interval to	52	52	20	NP	Res
3	survival was months in the	52	47	80	VP	Res
3	in the placebo group had	51	45	80	VP	Res
3	ratio for disease progression or	51	51	60	NP	Res
3	significant between group difference in	51	49	100	NP	Text
3	the primary outcome occurred in	51	51	60	VP	Res
3	group and in of patients	50	41	100	NP	Res
3	of those in the placebo	50	37	80	PP	Res
3	percentage points confidence interval to	50	50	40	NP	Res
3	as compared with of patients	49	43	100	PP	Res
3	by the log rank test	48	36	60	PP	Res
3	the primary end point was	48	44	80	VP	Res
3	in the combination therapy group	47	22	60	PP	Res
3	in the standard therapy group	47	20	80	PP	Res
3	there was no significant between	47	46	100	VP	Text
3	a total of patients in	46	45	100	NP	Res
3	as compared with percent of	46	40	100	PP	Res
3	group and months in the	46	42	100	NP	Res
3	serious adverse events occurred in	46	44	80	VP	Res
3	in of those in the	45	37	100	PP	Text
3	ml per minute per m	45	19	80	NP	Res
3	was significantly higher in the	45	44	80	VP	Text
3	adjusted odds ratio ci to	44	26	20	NP	Res
3	and in of the patients	44	38	100	PP	Res
3	of in the placebo group	44	33	80	PP	Res
3	were more common in the	44	42	100	VP	Text
3	p for the comparison of	43	21	60	NP	Res
3	primary end point occurred in	43	39	80	VP	Res
3	the modified intention to treat	43	35	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	a total of of patients[4]	42	40	100	NP	Res
3	adverse events occurred in of	42	39	80	VP	Res
3	ci to p and the	42	37	60	NP	Res
3	in the intervention group and	42	24	80	PP	Res
3	per person years in the	42	21	100	PP	Res
3	was significantly lower in the	42	41	80	VP	Text
3	as compared with of the	41	39	100	PP	Text
3	at a dose of mg	41	26	60	PP	Res
3	during a median follow up	41	41	80	PP	Res
3	group and of in the	41	35	100	NP	Res
3	the proportion of patients with	41	37	80	NP	Res
3	of the patients who underwent	40	39	80	PP	Res
3	the median follow up was	40	38	80	VP	Res
3	confidence interval ci to with	39	39	40	NP	Res
3	as compared with the placebo	38	36	80	PP	Res
3	confidence interval ci to for	38	38	40	NP	Res
3	in of the patients who	38	25	100	PP	Res
3	increase in the risk of	38	28	100	NP	Res
3	the percentage of patients with	38	33	80	NP	Res
3	and percent confidence interval to	37	30	60	NP	Res
3	group relative risk ci to	37	29	60	NP	Res
3	group relative risk confidence interval	37	37	40	NP	Res
3	in the placebo group relative	37	30	60	PP	Res
3	the rate of death from	37	33	100	NP	Res
3	were confidence interval ci to	37	37	40	VP	Res
3	as compared with the control	36	32	100	PP	Res
3	ci to in the placebo	36	29	60	NP	Res
3	of death from cardiovascular causes	36	35	80	PP	Res
3	or death ci to p	36	34	60	NP	Res
3	reduction in the risk of	36	25	80	NP	Res
3	vs relative risk ci to	36	16	40	PP	Res
3	adjusted hazard ratio ci to	35	22	20	NP	Res
3	person years of follow up	35	29	100	NP	Res
3	the rate of the primary	35	27	80	NP	Res
3	group hazard ratio in the	34	33	60	NP	Res
3	in the group assigned to	34	17	80	PP	Res
3	odds ratio confidence interval to	34	34	20	NP	Res
3	the placebo group in the	34	30	80	NP	Res
3	hazard ratio ci to and	33	29	40	NP	Res
3	in of participants in the	33	18	80	PP	Res
3	percent in the placebo group	33	30	80	NP	Res
3	progression or death confidence interval	33	33	40	NP	Res
3	underwent randomization were assigned to	33	33	40	VP	Res
3	in the risk of death	32	28	100	PP	Res
3	in the two study groups	32	31	100	PP	Res
3	median follow up period of	32	32	80	NP	Res
3	of the participants in the	32	22	80	PP	Res
3	the incidence of adverse events	32	30	60	NP	Res
3	were more likely to have	32	29	100	VP	Participant
3	in percent of the patients	31	28	100	PP	Res
3	group confidence interval to p	30	30	40	NP	Res
3	p for the comparison with	30	22	60	NP	Res
3	the hazard ratio for death	30	23	60	NP	Res
3	the presence or absence of	30	28	60	NP	Text
3	adjusted odds ratio percent confidence	29	20	20	NP	Res
3	adverse events was similar in	29	29	80	VP	Res
3	group and in of the	29	25	100	NP	Res
3	group than in the standard	29	20	100	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	in and of the patients	29	28	100	PP	Res
3	in the mg group and	29	20	80	PP	Res
3	months vs. months hazard ratio	29	23	40	NP	Res
3	patients in the rivaroxaban group	29	15	80	NP	Res
3	percent and percent respectively p	29	25	60	NP	Res
3	percent of patients in the	29	15	100	NP	Res
3	percent of those in the	29	22	100	NP	Res
3	reduction in the rate of	29	25	80	NP	Res
3	relative risk confidence interval to	29	29	40	NP	Res
3	the percentage of patients who	29	24	80	NP	Res
3	younger than years of age	29	26	100	Oth	Res
3	adverse events were reported in	28	28	80	VP	Res
3	patients who could be evaluated	28	24	80	NP	Res
3	the median age of the	28	28	80	NP	Res
3	a mean follow up of	27	27	100	NP	Res
3	among the patients in the	27	22	100	PP	Res
3	and of the patients in	27	26	100	PP	Res
3	differences in the rates of	27	26	100	NP	Res
3	incidence rate ratio ci to	27	12	40	NP	Res
3	patients were assigned to the	27	25	80	VP	Res
3	percent of the patients in	27	19	100	NP	Res
3	percent vs. percent p and	27	26	60	NP	Res
3	between the ages of and	26	17	100	PP	Res
3	group p for both comparisons	26	23	60	NP	Res
3	in the standard care group	26	12	100	PP	Res
3	months confidence interval ci to	26	26	40	NP	Res
3	odds ratio ci to and	26	19	40	NP	Res
3	percent percent and percent respectively	26	15	80	NP	Res
3	rate ratio ci to p	26	15	40	NP	Res
3	the median age was years	26	26	80	VP	Res
3	a primary end point event	25	24	80	NP	Res
3	adjusted odds ratio confidence interval	25	25	0	NP	Res
3	an intention to treat analysis	25	25	80	NP	Res
3	as compared with patients in	25	25	100	PP	Res
3	change from baseline in the	25	22	80	NP	Text
3	for death from any cause	25	24	100	PP	Res
3	in the control group vs	25	20	80	PP	Res
3	in the intensive therapy group	25	12	60	PP	Res
3	mmol per liter in the	25	11	60	NP	Res
3	months ci to in the	25	17	80	NP	Res
3	of adverse events were similar	25	25	80	VP	Res
3	outcome occurred in of patients	25	25	80	VP	Res
3	progression free survival was significantly	25	24	40	VP	Res
3	received at least one dose	25	23	80	VP	Res
3	reported in of the patients	25	25	100	VP	Res
3	the data and safety monitoring	25	25	60	NP	Res
3	the risk of death from	25	19	100	NP	Res
3	a difference of percentage points	24	17	80	NP	Res
3	a total of patients with	24	24	100	NP	Res
3	and were included in the	24	24	100	VP	Text
3	death percent confidence interval to	24	18	60	NP	Res
3	group and the control group	24	23	100	NP	Res
3	group percent confidence interval to	24	15	60	NP	Res
3	in of patients who received	24	17	100	PP	Res
3	in the rivaroxaban group and	24	12	80	PP	Res
3	in the vaccine group and	24	15	80	PP	Res
3	no significant differences in the	24	22	100	NP	Text
3	of the patients treated with	24	19	100	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	and of the patients respectively	23	20	80	NP	Res
3	as compared with those who	23	17	100	PP	Text
3	associated with an increase in	23	23	100	VP	Text
3	during a mean follow up	23	23	100	PP	Res
3	end point occurred in patients	23	19	100	VP	Res
3	events occurred in of the	23	23	100	VP	Res
3	follow up of years the	23	23	100	NP	Res
3	followed for a median of	23	23	80	VP	Res
3	in the high dose group	23	12	80	PP	Res
3	of percentage points ci to	23	21	60	PP	Res
3	of the patients assigned to	23	16	80	PP	Res
3	outcome occurred in patients in	23	22	80	VP	Res
3	the weight in kilograms divided	23	23	60	NP	Res
3	therapy group than in the	23	17	80	NP	Res
3	there were deaths in the	23	23	100	VP	Res
3	vs odds ratio ci to	23	18	20	PP	Res
3	with a median follow up	23	23	80	PP	Res
3	with of patients in the	23	22	100	PP	Res
3	a dose of mg per	22	14	60	NP	Res
3	as compared with percent in	22	21	100	PP	Res
3	as compared with placebo was	22	19	80	VP	Res
3	between the groups in the	22	21	100	PP	Res
3	confidence interval to p for	22	17	40	NP	Res
3	difference confidence interval ci to	22	22	40	NP	Res
3	difference in the rate of	22	22	100	NP	Res
3	events occurred in of patients	22	20	100	NP	Res
3	follow up period of years	22	22	100	NP	Res
3	group hazard ratio for disease	22	22	60	NP	Res
3	had a sustained virologic response	22	15	60	VP	Res
3	in the per protocol analysis	22	20	80	PP	Res
3	in the usual care group	22	13	80	PP	Res
3	months in the placebo group	22	20	80	NP	Res
3	patients were included in the	22	22	100	VP	Res
3	percentage of patients with a	22	19	80	NP	Res
3	placebo hazard ratio ci to	22	20	20	NP	Res
3	rates of serious adverse events	22	22	80	NP	Res
3	ratio for death in the	22	22	80	NP	Res
3	the ages of and years	22	16	100	NP	Res
3	the objective response rate was	22	22	80	VP	Res
3	the upper limit of the	22	20	80	NP	Res
3	there was no evidence of	22	22	100	VP	Text
3	vs in the placebo group	22	20	60	PP	Res
3	was ci to in the	22	19	80	VP	Res
3	was significantly longer in the	22	22	80	VP	Text
3	with respect to the primary	22	22	80	PP	Res
3	a total of patients received	21	21	100	NP	Res
3	at years was in the	21	15	100	PP	Res
3	end point occurred in of	21	21	100	VP	Res
3	free survival was significantly longer	21	20	60	VP	Res
3	hazard ratio for death p	21	20	40	NP	Res
3	in the per protocol population	21	16	80	PP	Res
3	in the placebo group of	21	20	80	PP	Res
3	median age of the patients	21	21	80	NP	Res
3	occurred in and of the	21	20	100	VP	Res
3	of the patients who were	21	19	100	PP	Res
3	on the basis of the	21	20	100	PP	Text
3	participants in the placebo group	21	16	60	NP	Res
3	patients in the intervention group	21	11	80	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	patients were assigned to receive	21	19	80	VP	Res
3	rate of sustained virologic response	21	16	60	NP	Res
3	serious adverse events were reported	21	21	80	VP	Res
3	significant difference between the two	21	21	100	NP	Text
3	survival was months ci to	21	18	60	VP	Res
3	the group given mg of	21	10	80	NP	Res
3	the overall response rate was	21	20	80	VP	Res
3	there were no serious adverse	21	21	80	VP	Res
3	was associated with a significantly	21	20	80	VP	Text
3	were included in the analysis	21	21	100	VP	Res
3	were percent confidence interval to	21	17	60	VP	Res
3	were reported in of the	21	21	100	VP	Res
3	a percent reduction in the	20	14	80	NP	Res
3	adjusted hazard ratio confidence interval	20	20	0	NP	Res
3	as compared with months in	20	17	100	PP	Res
3	at the end of the	20	18	100	PP	Res
3	ci to p and a	20	19	60	NP	Res
3	control group hazard ratio for	20	18	60	NP	Res
3	difference between the two groups	20	20	100	NP	Text
3	for the primary end point	20	19	80	PP	Res
3	group odds ratio ci to	20	15	40	NP	Res
3	group odds ratio confidence interval	20	20	20	NP	Res
3	in of infants in the	20	8	80	PP	Res
3	in of patients assigned to	20	15	80	PP	Res
3	in the intensive care unit	20	20	80	PP	Res
3	interquartile range to in the	20	10	80	NP	Res
3	less than copies per milliliter	20	15	60	Oth	Res
3	occurred in of patients receiving	20	19	100	VP	Res
3	the mean age of the	20	20	100	NP	Res
3	the median duration of response	20	20	60	NP	Res
3	therapy was associated with a	20	18	80	VP	Res
3	to years of age and	20	18	100	PP	Res
3	were followed for a median	20	20	80	VP	Res
3	were more frequent in the	20	19	80	VP	Text
3	were randomly assigned to the	20	20	60	VP	Res
3	were to years of age	20	13	100	VP	Res
3	years of follow up the	20	18	100	NP	Res
3	a total of participants were	19	19	80	NP	Res
3	confidence interval to in the	19	12	60	NP	Res
3	confidence interval to percent and	19	16	60	NP	Res
3	did not differ significantly among	19	18	60	VP	Text
3	follow up period of months	19	19	100	NP	Res
3	forced expiratory volume in second	19	19	60	NP	Res
3	group as compared with percent	19	18	100	NP	Res
3	group percent vs. percent p	19	16	60	NP	Res
3	hazard ratio for death from	19	16	60	NP	Res
3	in of the patients and	19	18	100	PP	Res
3	modified intention to treat population	19	15	60	NP	Res
3	of adverse events was similar	19	19	80	VP	Res
3	of percentage points confidence interval	19	19	40	PP	Res
3	of years of follow up	19	18	100	PP	Res
3	p for the comparison between	19	13	60	NP	Res
3	patients who underwent randomization were	19	19	60	VP	Res
3	per deciliter μmol per liter	19	10	40	PP	Res
3	primary outcome occurred in patients	19	19	60	VP	Res
3	rate of progression free survival	19	16	60	NP	Res
3	relative risk ci to and	19	17	60	NP	Res
3	survival was months confidence interval	19	19	40	VP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	vs hazard ratio confidence interval	19	19	0	PP	Res
3	who were randomly assigned to	19	15	60	VP	Res
3	a median of years of	18	17	80	NP	Res
3	absolute risk difference percentage points	18	15	60	NP	Res
3	and percent of those in	18	17	100	NP	Res
3	and the placebo group in	18	16	80	NP	Res
3	associated with a reduction in	18	17	80	VP	Text
3	ci to p by the	18	12	60	NP	Res
3	confidence interval for the difference	18	15	60	NP	Res
3	confidence interval to p by	18	16	40	NP	Res
3	duration of follow up was	18	18	80	NP	Res
3	end point event occurred in	18	17	100	VP	Res
3	in the control group had	18	17	100	PP	Res
3	in the group given mg	18	8	80	PP	Res
3	in the sirolimus stent group	18	8	60	PP	Res
3	intervention group and in the	18	15	80	NP	Res
3	level was mg per deciliter	18	15	60	VP	Res
3	months hazard ratio for death	18	18	60	NP	Res
3	occurred more frequently in the	18	17	80	VP	Res
3	of ci to in the	18	11	80	PP	Res
3	on the modified rankin scale	18	15	40	PP	Res
3	p and the rate of	18	16	80	NP	Res
3	p for both comparisons with	18	16	60	NP	Res
3	percent confidence interval to as	18	15	60	NP	Res
3	rates of adverse events were	18	18	80	NP	Res
3	respectively relative risk ci to	18	15	40	NP	Res
3	significant differences between the two	18	18	100	NP	Text
3	the patients in the placebo	18	16	80	NP	Res
3	therapy group and in the	18	13	80	NP	Res
3	to with higher scores indicating	18	18	100	PP	Text
3	was associated with a higher	18	18	100	VP	Text
3	was associated with a reduction	18	18	80	VP	Text
3	was associated with a significant	18	18	100	VP	Text
3	with a reduction in the	18	17	80	PP	Text
3	and death from any cause	17	14	100	NP	Res
3	did not differ between the	17	17	80	VP	Text
3	group and in the mg	17	12	80	NP	Res
3	group and in the standard	17	10	100	NP	Res
3	group hazard ratio for progression	17	17	40	NP	Res
3	hazard ratio for death with	17	15	60	NP	Res
3	in of the patients with	17	16	100	PP	Res
3	in the placebo group percent	17	16	80	PP	Res
3	intervention group than in the	17	14	80	NP	Res
3	mean follow up of years	17	17	100	NP	Res
3	of patients who underwent randomization	17	17	60	PP	Res
3	of those who received placebo	17	15	80	PP	Res
3	patients were enrolled in the	17	15	80	VP	Res
3	per kilogram of body weight	17	17	80	PP	Res
3	per person years ci to	17	8	80	PP	Res
3	reached the primary end point	17	16	80	VP	Res
3	the adjusted hazard ratio for	17	13	40	NP	Res
3	the time to the first	17	13	100	NP	Res
3	to the primary end point	17	16	80	PP	Res
3	was associated with a lower	17	17	100	VP	Text
3	were percent percent and percent	17	11	100	VP	Res
3	were significantly higher in the	17	17	80	VP	Text
3	adverse events were more common	16	15	80	VP	Res
3	and to the control group	16	16	100	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	as compared with ci to	16	14	80	PP	Res
3	as compared with patients with	16	15	100	PP	Res
3	associated with the risk of	16	14	100	VP	Res
3	characteristics were similar in the	16	16	80	VP	Text
3	confidence interval ci to or	16	16	40	NP	Res
3	did not differ significantly from	16	16	60	VP	Text
3	differences between the groups in	16	16	100	NP	Text
3	grade or adverse events were	16	16	60	VP	Res
3	group as compared with months	16	14	100	NP	Res
3	health related quality of life	16	16	100	NP	Res
3	in an intention to treat	16	16	80	PP	Res
3	in the aspirin group and	16	9	80	PP	Res
3	in the intervention group than	16	13	80	PP	Res
3	in the rate of death	16	14	100	PP	Res
3	in the two groups respectively	16	15	80	PP	Res
3	in the zoledronic acid group	16	5	60	PP	Res
3	incidence of adverse events was	16	16	60	NP	Res
3	mean age of the patients	16	16	100	NP	Res
3	median duration of response was	16	16	60	NP	Res
3	occurred in patients assigned to	16	13	80	VP	Res
3	of grade or higher were	16	16	80	VP	Res
3	or death from any cause	16	14	100	NP	Res
3	patients in the rituximab group	16	7	80	NP	Res
3	per person years hazard ratio	16	6	60	PP	Res
3	percent confidence interval to respectively	16	14	40	NP	Res
3	percent in the control group	16	12	100	NP	Res
3	percent vs. percent relative risk	16	12	60	NP	Res
3	percentage points ci to and	16	13	60	NP	Res
3	significant difference between the groups	16	12	100	NP	Res
3	the group receiving mg of	16	9	80	NP	Res
3	the hazard ratio for the	16	14	60	NP	Res
3	the incidence of the primary	16	15	60	NP	Res
3	the median time to the	16	14	80	NP	Res
3	the rate of adverse events	16	16	80	NP	Res
3	those years of age or	16	14	100	NP	Res
3	to in the control group	16	15	100	PP	Res
3	to percent confidence interval to	16	10	60	PP	Res
3	was not associated with a	16	16	100	VP	Text
3	was significantly greater in the	16	15	80	VP	Text
3	were percent and percent respectively	16	13	80	VP	Res
3	were years of age or	16	16	100	VP	Res
3	years hazard ratio ci to	16	9	40	NP	Res
3	a significant reduction in the	15	13	80	NP	Res
3	a total of patients had	15	15	100	VP	Res
3	a total of women were	15	15	100	VP	Res
3	among the patients who received	15	13	100	NP	Res
3	among those who received placebo	15	14	80	NP	Res
3	and months respectively hazard ratio	15	11	40	NP	Res
3	as compared with those with	15	15	100	PP	Res
3	blood pressure was mm hg	15	13	40	VP	Res
3	body mass index the weight	15	15	60	NP	Res
3	bone mineral density at the	15	10	40	NP	Res
3	ci to p for superiority	15	14	40	NP	Res
3	death confidence interval to p	15	15	40	NP	Res
3	for a median of years	15	15	80	PP	Res
3	free survival was months with	15	14	80	VP	Res
3	from any cause hazard ratio	15	13	60	PP	Res
3	group p for the comparison	15	12	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	group vs. in the placebo	15	14	60	NP	Res
3	in and of patients respectively	15	14	80	PP	Res
3	in the control group and	15	13	100	PP	Res
3	in the pci group and	15	10	80	PP	Res
3	in the placebo group difference	15	14	80	PP	Res
3	in the placebo group for	15	12	80	PP	Res
3	in the placebo group were	15	15	80	VP	Res
3	in the surgery group and	15	8	80	PP	Res
3	in the two groups in	15	14	100	PP	Res
3	incidence of serious adverse events	15	15	60	NP	Res
3	less than mg per deciliter	15	7	60	Oth	Res
3	median follow up was years	15	15	80	VP	Res
3	mg group and in the	15	12	80	NP	Res
3	months in the control group	15	15	100	NP	Res
3	occurred in of participants in	15	11	80	VP	Res
3	occurred in of patients who	15	15	100	VP	Res
3	occurred in of the participants	15	12	80	VP	Res
3	of progression free survival was	15	13	60	VP	Res
3	of the women in the	15	10	100	PP	Res
3	over a median follow up	15	15	80	PP	Res
3	patients who had a response	15	14	100	NP	Res
3	per minute per m of	15	13	100	PP	Res
3	percent as compared with percent	15	15	100	NP	Res
3	randomization were assigned to the	15	15	60	VP	Res
3	respectively hazard ratio confidence interval	15	15	0	NP	Res
3	serious adverse events were similar	15	15	80	VP	Res
3	than among those who received	15	15	100	Oth	Res
3	the placebo group and the	15	14	80	NP	Res
3	the primary efficacy end point	15	15	60	NP	Res
3	therapy group as compared with	15	14	80	NP	Res
3	vaccine group and in the	15	11	80	NP	Res
3	was associated with an increase	15	15	100	VP	Text
3	was confidence interval to p	15	15	40	VP	Res
3	was months and months respectively	15	10	80	VP	Res
3	was observed in of patients	15	14	80	VP	Res
3	with an increase in the	15	15	100	PP	Res
3	with the use of the	15	14	100	PP	Res
3	years of age or younger	15	13	100	NP	Res
3	a total of participants underwent	14	14	60	VP	Res
3	adverse events occurred in patients	14	13	80	VP	Res
3	an increased risk of death	14	13	100	NP	Res
3	and in the mg group	14	11	80	PP	Res
3	between the two treatment groups	14	14	100	PP	Res
3	by confidence interval ci to	14	14	40	PP	Res
3	cause hazard ratio ci to	14	11	40	NP	Res
3	ci to for death from	14	7	80	NP	Res
3	ci to in the group	14	6	80	NP	Res
3	ci to p and in	14	14	60	NP	Res
3	ci to p for the	14	9	60	NP	Res
3	ci to p hazard ratio	14	11	20	NP	Res
3	death from any cause or	14	13	100	NP	Res
3	death from any cause was	14	14	100	VP	Res
3	difference in the incidence of	14	14	80	NP	Res
3	group p for all comparisons	14	11	60	NP	Res
3	groups did not differ significantly	14	14	60	VP	Text
3	in the change from baseline	14	14	80	PP	Res
3	in the control group in	14	13	100	PP	Res
3	in the endovascular repair group	14	5	60	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	in the group receiving mg	14	6	80	PP	Res
3	in the low dose group	14	6	80	PP	Res
3	in the placebo group as	14	14	80	PP	Res
3	in the subgroup of patients	14	11	80	PP	Res
3	in the treatment group and	14	5	100	PP	Res
3	in the two treatment groups	14	14	100	PP	Res
3	length of stay in the	14	12	80	NP	Res
3	major bleeding occurred in patients	14	11	80	VP	Res
3	no serious adverse events were	14	14	80	VP	Res
3	of patients vs. of patients	14	11	80	PP	Res
3	patients confidence interval ci to	14	14	40	NP	Res
3	patients years of age or	14	10	100	NP	Res
3	percent confidence interval for the	14	11	60	NP	Res
3	primary outcome event occurred in	14	13	60	VP	Res
3	rates of sustained virologic response	14	13	60	NP	Res
3	ratio ci to p but	14	14	40	NP	Res
3	ratio ci to p or	14	14	40	NP	Res
3	response rate was ci to	14	12	80	VP	Res
3	the change from baseline in	14	13	80	NP	Res
3	the control group in the	14	13	100	NP	Res
3	the most common adverse event	14	14	80	NP	Res
3	the proportion of patients who	14	14	80	NP	Res
3	the rates of adverse events	14	14	80	NP	Res
3	the trial was stopped early	14	14	80	VP	Res
3	there were also no significant	14	14	100	VP	Text
3	times the upper limit of	14	13	80	NP	Res
3	treatment related adverse events of	14	13	80	NP	Res
3	was achieved in of patients	14	10	100	VP	Res
3	was of patients in the	14	12	100	VP	Res
3	a median of days after	13	11	80	NP	Res
3	and months ci to in	13	11	80	NP	Res
3	and of women in the	13	9	100	NP	Res
3	as compared with those in	13	12	100	PP	Res
3	at the time of the	13	13	100	PP	Res
3	cases per person years in	13	7	100	NP	Res
3	ci to p as was	13	13	60	VP	Res
3	ci to p for trend	13	7	40	NP	Res
3	common grade or adverse events	13	13	60	NP	Res
3	confidence interval to p as	13	13	40	NP	Res
3	days confidence interval ci to	13	13	40	NP	Res
3	death from coronary heart disease	13	9	80	NP	Res
3	difference between the groups in	13	11	100	NP	Res
3	difference in risk percentage points	13	8	80	NP	Res
3	dose of mg per kilogram	13	9	40	NP	Res
3	duration of progression free survival	13	12	40	NP	Res
3	for a median of months	13	13	80	PP	Res
3	free survival was months and	13	12	80	VP	Res
3	group absolute difference percentage points	13	10	60	NP	Res
3	group and days in the	13	13	100	NP	Res
3	group in the mg group	13	8	80	NP	Res
3	group risk difference percentage points	13	10	80	NP	Res
3	hazard ratio for death or	13	11	60	NP	Res
3	hazard ratio was ci to	13	12	40	VP	Res
3	in of the patients receiving	13	10	100	PP	Res
3	in of those who received	13	8	100	PP	Res
3	in the combined therapy group	13	5	60	PP	Res
3	in the conventional therapy group	13	8	60	PP	Res
3	in the group receiving the	13	8	100	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	in the mg tofacitinib group	13	5	60	PP	Res
3	in the standard treatment group	13	5	100	PP	Res
3	observed in of the patients	13	13	80	VP	Res
3	of log copies per milliliter	13	9	40	PP	Res
3	of patients who had a	13	12	100	PP	Res
3	of serious adverse events was	13	13	80	VP	Res
3	of the composite end point	13	10	80	PP	Res
3	of the patients who had	13	13	100	PP	Res
3	of the primary outcome was	13	13	60	VP	Res
3	patients in the aspirin group	13	8	80	NP	Res
3	percentage of patients who were	13	11	80	NP	Res
3	placebo p for both comparisons	13	11	40	NP	Res
3	risk ratio ci to p	13	12	40	NP	Res
3	significantly more patients in the	13	12	80	NP	Res
3	survival hazard ratio for death	13	12	40	NP	Res
3	survival was months and months	13	9	80	VP	Res
3	that in the placebo group	13	12	80	NP	Res
3	the area under the curve	13	11	80	NP	Res
3	the rate of grade or	13	10	80	NP	Res
3	the two groups were similar	13	13	100	VP	Text
3	venous thromboembolism occurred in of	13	11	60	VP	Res
3	was also associated with a	13	13	100	VP	Text
3	was cells per cubic millimeter	13	13	60	VP	Res
3	was ci to and the	13	11	80	VP	Res
3	was more frequent in the	13	12	80	VP	Res
3	was observed in of the	13	13	80	VP	Res
3	were more likely to be	13	13	100	VP	Participant
3	with a reduced risk of	13	10	100	PP	Res
3	adjusted hazard ratio for death	12	7	40	NP	Res
3	among the patients who were	12	11	100	VP	Res
3	and p for the comparison	12	8	60	NP	Res
3	and percentage points ci to	12	9	60	NP	Res
3	and serious adverse events were	12	12	80	VP	Res
3	and the hazard ratio for	12	12	60	NP	Res
3	and were more likely to	12	12	100	VP	Participant
3	ci to and in the	12	10	80	NP	Res
3	ci to hazard ratio for	12	11	40	NP	Res
3	ci to p and death	12	10	60	NP	Res
3	ci to p for both	12	12	60	NP	Res
3	difference percent percent confidence interval	12	9	60	NP	Res
3	during the follow up period	12	12	100	PP	Res
3	end point of death from	12	11	100	NP	Res
3	end point was in the	12	10	100	VP	Res
3	from any cause ci to	12	12	80	PP	Res
3	group p and in the	12	12	80	NP	Res
3	had a higher rate of	12	12	100	VP	Text
3	had occurred in of patients	12	10	100	VP	Res
3	in the control group percent	12	9	100	PP	Res
3	in the ldl cholesterol level	12	10	60	PP	Res
3	in the placebo group died	12	11	80	PP	Res
3	in the placebo group with	12	11	80	PP	Res
3	in the proportion of patients	12	11	80	PP	Res
3	not associated with an increased	12	12	100	VP	Text
3	occurred in participants in the	12	12	80	VP	Res
3	of follow up was years	12	12	100	VP	Res
3	of in the control group	12	11	100	PP	Res
3	of the children in the	12	8	100	PP	Res
3	of the infants in the	12	6	80	PP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	of the patients in group	12	5	100	PP	Res
3	other adverse events were similar	12	12	80	VP	Res
3	outcome event occurred in of	12	11	80	VP	Res
3	patients in the surgery group	12	9	80	NP	Res
3	patients vs. patients hazard ratio	12	9	40	NP	Res
3	patients who did not have	12	12	100	NP	Res
3	percentage points ci to for	12	6	60	NP	Res
3	rate of overall survival was	12	12	60	NP	Res
3	response rate was in the	12	12	100	VP	Res
3	risk of death from any	12	11	100	NP	Res
3	serious adverse events in the	12	12	80	NP	Res
3	the median follow up period	12	12	80	NP	Res
3	the placebo group had a	12	11	80	VP	Res
3	the placebo group vs. p	12	12	40	NP	Res
3	the primary end point of	12	12	80	NP	Res
3	the risk of death was	12	11	100	VP	Res
3	there was a trend toward	12	12	80	VP	Text
3	to p as was the	12	12	80	PP	Res
3	vs hazard ratio for death	12	11	40	PP	Res
3	vs in the control group	12	9	80	PP	Res
3	was in the placebo group	12	9	80	VP	Res
3	year in the placebo group	12	10	80	NP	Res
3	a higher percentage of patients	11	10	80	NP	Res
3	a percent increase in the	11	5	100	NP	Res
3	a relative reduction of in	11	11	60	NP	Res
3	a significantly increased risk of	11	10	80	NP	Res
3	a total of children were	11	11	100	VP	Res
3	among patients who did not	11	9	100	PP	Text
3	among patients who had a	11	9	100	PP	Text
3	among patients years of age	11	8	100	PP	Res
3	and ci to for the	11	9	80	NP	Res
3	as compared with with placebo[5]	11	9	80	PP	Res
3	at the end of treatment	11	9	100	PP	Res
3	between group difference percentage points	11	10	80	PP	Res
3	between the two study groups	11	11	100	PP	Res
3	death hazard ratio ci to	11	11	40	NP	Res
3	did not differ significantly in	11	11	60	VP	Text
3	event occurred in of patients	11	10	100	VP	Res
3	group and months ci to	11	9	80	NP	Res
3	group had died relative risk	11	9	80	VP	Res
3	groups in the rates of	11	11	100	NP	Res
3	had an increased risk of	11	11	100	VP	Res
3	had no significant effect on	11	11	100	VP	Text
3	hazard ratio for disease recurrence	11	9	40	NP	Res
3	in the active treatment group	11	5	80	PP	Res
3	in the apixaban group and	11	6	80	PP	Res
3	in the control group adjusted	11	9	80	PP	Res
3	in the hypothermia group and	11	7	80	PP	Res
3	in the paclitaxel stent group	11	5	60	PP	Res
3	in the placebo group incidence	11	8	60	PP	Res
3	in the t pr group	11	5	80	PP	Res
3	in the watchful waiting group	11	6	80	PP	Res
3	liter percent confidence interval to	11	6	40	NP	Res
3	mean change from baseline in	11	10	80	NP	Res
3	mg per deciliter in the	11	5	60	NP	Res
3	no significant differences among the	11	11	100	NP	Text
3	no significant differences were observed	11	11	80	VP	Text
3	occurred in and of patients	11	9	100	VP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	occurred in patients who received	11	10	100	VP	Res
3	occurred in percent of the	11	11	100	VP	Res
3	of grade or higher occurred	11	11	80	PP	Res
3	of less than mg per	11	7	80	PP	Res
3	patients had at least one	11	11	100	VP	Res
3	patients in the apixaban group	11	6	80	NP	Res
3	patients in the pci group	11	6	80	NP	Res
3	patients who did not receive	11	9	100	NP	Res
3	percent confidence interval to or	11	10	60	NP	Res
3	percent vs. percent p or	11	11	60	NP	Res
3	placebo hazard ratio confidence interval	11	11	0	NP	Res
3	progression free survival rate was	11	10	60	VP	Res
3	rate of serious adverse events	11	11	80	NP	Res
3	rate of the primary outcome	11	11	60	NP	Res
3	reported in of patients in	11	11	100	VP	Res
3	risk of death from cardiovascular	11	11	80	NP	Res
3	the end of the study	11	11	100	NP	Res
3	the most frequent adverse events	11	11	60	NP	Res
3	the patients randomly assigned to	11	8	60	NP	Res
3	the placebo group p and	11	11	60	NP	Res
3	the primary composite end point	11	11	60	NP	Res
3	the rate of freedom from	11	9	80	NP	Res
3	the rates of the primary	11	11	80	NP	Res
3	to hazard ratio ci to	11	7	40	PP	Res
3	was associated with a percent	11	10	100	VP	Text
3	was more common in the	11	11	100	VP	Text
3	were ci to ci to	11	11	60	VP	Res
3	were lost to follow up	11	10	100	VP	Res
3	were randomly assigned to a	11	11	60	VP	Res
3	who did not have a	11	11	100	VP	Res
3	who received at least one	11	9	100	VP	Res
3	who received placebo p for	11	7	60	VP	Res
3	women in the placebo group	11	9	80	NP	Res
3	years confidence interval ci to	11	11	40	NP	Res
3	a glycated hemoglobin level of	10	7	60	NP	Res
3	a significant increase in the	10	9	100	NP	Res
3	a significantly lower risk of	10	9	80	NP	Res
3	a total of cases of	10	10	100	NP	Res
3	a total of infants were	10	10	80	VP	Res
3	a total of participants in	10	10	80	NP	Res
3	an increase in the risk	10	9	100	NP	Res
3	and events per person years	10	9	100	NP	Res
3	and in the group receiving	10	6	100	PP	Res
3	and of participants in the	10	7	80	NP	Res
3	and percent in the group	10	9	100	NP	Res
3	and percent respectively in the	10	6	80	NP	Res
3	as compared with percent among	10	10	100	PP	Res
3	as compared with placebo p	10	7	60	PP	Res
3	at months the rate of	10	10	100	PP	Res
3	at to years of age	10	7	100	PP	Res
3	at years the rate of	10	10	100	PP	Res
3	bleeding occurred in of patients	10	10	80	VP	Res
3	by the end of the	10	10	100	PP	Res
3	ci to among patients with	10	8	80	NP	Res
3	ci to in the per	10	7	80	NP	Res
3	ci to p and for	10	8	60	NP	Res
3	confidence interval ci to not	10	10	40	NP	Res
3	control group than in the	10	8	100	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	data and safety monitoring committee	10	10	40	NP	Res
3	days of patients in the	10	8	100	NP	Res
3	death from any cause occurred	10	10	100	VP	Res
3	developed in of the patients	10	10	100	VP	Res
3	developed in patients in the	10	10	100	VP	Res
3	died as compared with of	10	9	100	VP	Res
3	differences in the rate of	10	9	100	NP	Res
3	for progression or death ci	10	8	60	PP	Res
3	for the comparison between the	10	6	80	PP	Text
3	free survival and overall survival	10	9	40	NP	Res
3	free survival hazard ratio for	10	10	40	NP	Res
3	free survival was in the	10	9	80	VP	Res
3	from any cause occurred in	10	10	100	PP	Res
3	group and in the medical	10	10	80	NP	Res
3	group and in the surgery	10	9	80	NP	Res
3	group in the intention to	10	10	80	NP	Res
3	group of patients vs. of	10	10	80	NP	Res
3	group rate ratio ci to	10	10	60	NP	Res
3	had a lower risk of	10	10	100	VP	Res
3	had received a median of	10	10	80	VP	Res
3	hazard ratios for death from	10	8	60	NP	Res
3	in the budesonide formoterol group	10	5	60	PP	Res
3	in the group that underwent	10	6	80	PP	Res
3	in the placebo group adjusted	10	5	60	PP	Res
3	in the rate of the	10	10	100	PP	Res
3	in the rates of death	10	10	100	PP	Res
3	in the surgery group p	10	8	60	PP	Res
3	in the two groups were	10	10	100	PP	Res
3	in the two groups with	10	10	100	PP	Res
3	increase in the incidence of	10	8	80	NP	Res
3	increased by a factor of	10	9	100	VP	Res
3	increased from in to in	10	9	100	VP	Text
3	independent data and safety monitoring	10	10	40	NP	Res
3	intention to treat analysis of	10	10	80	NP	Res
3	intervention group as compared with	10	8	80	NP	Res
3	kaplan meier estimates of the	10	10	60	NP	Res
3	level of mg per deciliter	10	7	60	NP	Res
3	major bleeding occurred in of	10	10	80	VP	Res
3	mg group as compared with	10	6	80	NP	Res
3	occurred in patients and patients	10	9	100	VP	Res
3	of the patients had a	10	10	100	PP	Res
3	of the patients in each	10	10	100	PP	Res
3	older than years of age	10	10	100	Oth	Res
3	outcomes did not differ significantly	10	10	40	VP	Res
3	patients in the enoxaparin group	10	7	80	NP	Res
3	patients in the respective groups	10	10	80	NP	Res
3	percent confidence interval to with	10	6	60	NP	Res
3	percent in the group given	10	5	100	NP	Res
3	progression free survival in the	10	10	60	NP	Res
3	progression or death in the	10	10	80	NP	Res
3	rate of disease free survival	10	9	80	NP	Res
3	reduction in the number of	10	8	80	NP	Res
3	risk of death hazard ratio	10	10	60	NP	Res
3	significant between group differences were	10	10	100	VP	Res
3	standard therapy group hazard ratio	10	6	40	NP	Res
3	surgery group and in the	10	8	80	NP	Res
3	survival was similar in the	10	10	80	VP	Res
3	sustained virologic response at weeks	10	5	60	NP	Res

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
3	than among those receiving placebo	10	10	80	Oth	Res
3	than in the group that	10	10	100	Oth	Res
3	than with placebo vs. p	10	8	40	Oth	Res
3	the between group difference in	10	9	100	NP	Res
3	the control group p and	10	9	80	NP	Res
3	the control group vs. p	10	9	60	NP	Res
3	the group that received the	10	7	100	NP	Res
3	the mg group and the	10	5	80	NP	Res
3	the most common grade or	10	10	80	NP	Res
3	the patients who had a	10	10	100	NP	Res
3	the patients who were enrolled	10	10	80	NP	Res
3	the placebo group and respectively	10	9	60	NP	Res
3	the rate of death was	10	10	100	VP	Res
3	the relative risk of death	10	9	80	NP	Res
3	the results were similar in	10	10	100	VP	Text
3	the subgroup of patients with	10	10	80	NP	Res
3	the two groups did not	10	10	100	VP	Text
3	those who received placebo p	10	8	60	NP	Res
3	treatment group than in the	10	8	100	NP	Res
3	vs p as was the	10	10	60	PP	Res
3	was days interquartile range to	10	10	80	VP	Res
3	was reported in of patients	10	9	100	VP	Res
3	was similar to that of	10	10	100	VP	Text
3	were enrolled in the study	10	10	80	VP	Res
3	were included in the analyses	10	10	100	VP	Res
3	were similar with respect to	10	10	100	VP	Text
3	women to years of age	10	5	100	NP	Res
3	women were randomly assigned to	10	10	60	VP	Res
4	associated with an increased risk	59	56	100	VP	Res
4	with an increased risk of	58	55	100	PP	Res
4	did not result in a	38	38	100	VP	Text
4	at a dose of mg	31	27	60	PP	Res
4	was associated with an increased	30	30	100	VP	Res
4	of death from any cause	27	24	100	PP	Res
4	there was no significant difference	27	27	100	VP	Text
4	than among those who received	26	26	100	Oth	Res
4	out of hospital cardiac arrest	23	22	60	NP	Res
4	was associated with a higher	23	23	100	VP	Res
4	did not differ significantly between	22	22	60	VP	Text
4	was not associated with a	22	22	100	VP	Text
4	did not significantly reduce the	21	21	80	VP	Text
4	a significantly lower rate of	20	20	80	NP	Text
4	years of age or older	20	18	100	NP	Res
4	is associated with an increased	19	19	100	VP	Text
4	reduction in the rate of	19	19	80	NP	Text
4	was associated with a lower	19	18	100	VP	Text
4	non small cell lung cancer	18	18	40	NP	Res
4	rates of sustained virologic response	18	18	60	NP	Res
4	in a lower incidence of	17	17	80	PP	Text
4	the rate of death from	17	17	100	NP	Res
4	was associated with a significantly	17	17	80	VP	Text
4	a significantly lower risk of	16	16	80	NP	Res
4	associated with a reduction in	15	15	80	VP	Text
4	lower among those who received	15	15	100	Oth	Res
4	not reduce the rate of	15	15	100	VP	Text
4	patients with moderate to severe	15	15	60	NP	Res
4	resulted in a significantly lower	15	15	80	VP	Text
4	and was associated with a	14	14	100	VP	Text

Table A1. *Cont.*

Move	Bundles	Freq	Range	1st N	Structure	Function
4	composite end point of death	14	14	80	NP	Res
4	free survival among patients with	14	14	80	NP	Res
4	in patients with type diabetes	14	14	80	PP	Res
4	in the rate of death	14	13	100	PP	Res
4	of death from cardiovascular causes	14	14	80	PP	Res
4	patients with relapsed or refractory	14	14	60	NP	Res
4	than those who received placebo	14	14	80	Oth	Res
4	an increased risk of death	13	13	100	NP	Res
4	free survival and overall survival	13	13	40	NP	Res
4	in patients with atrial fibrillation	13	13	60	PP	Res
4	not reduce the risk of	13	13	100	VP	Res
4	patients with hcv genotype infection	13	12	40	NP	Res
4	a significantly higher rate of	12	12	80	NP	Text
4	associated with an increase in	12	12	100	VP	Text
4	longer progression free survival than	12	12	60	NP	Res
4	with a lower rate of	12	12	100	PP	Text
4	with a lower risk of	12	11	100	PP	Res
4	among patients with type diabetes	11	11	80	NP	Res
4	had no significant effect on	11	11	100	VP	Res
4	in a lower rate of	11	10	100	PP	Text
4	in this trial involving patients	11	11	80	PP	Res
4	increase in the rate of	11	10	100	NP	Text
4	overall survival among patients with	11	11	60	NP	Res
4	than in the placebo group	11	11	80	Oth	Res
4	was associated with a significant	11	11	100	VP	Text
4	increase in the risk of	10	8	100	NP	Res
4	no significant difference in the	10	10	100	NP	Text
4	patients with acute ischemic stroke	10	10	40	NP	Res
4	patients with chronic kidney disease	10	10	60	NP	Res
4	the rate of the composite	10	10	80	NP	Res
4	the risk of death from	10	9	100	NP	Res
4	with a higher rate of	10	9	100	PP	Text

Notes

1 Although translations can sometimes be criticized for not perfectly reflecting "the content" of the original source texts or texts originally written in the target language (Rees 2022, p. 395), the Japanese translations of the journal's abstracts, shown in the regional edition of the journal as well as on the website of the licensed publisher (Nankodo 2024), are noted for accurately reflecting the original texts. Noto (2015) specifically highlighted their effectiveness in conveying essential information to the domestic disciplinary readers, ensuring that critical medical insights remain accessible and reliable.

2 When measured in the Google Colaboratory's Python environment, the English-language abstract texts comprised 1,201,780 words. However, using AntConc for quantification, the count was 1,209,851 words with punctuation included and 1,148,583 words without punctuation. For the purposes of this study, we adopt the more conservative word count of 1,148,583, excluding punctuation, to describe the size of our corpus.

3 The bundle "a total of of the" was extracted from fragments such as "a total of 444 of the 490 children" because punctuation and numerals were excluded when extracting bundles, following the methodology of a previous study (Durrant 2017, p. 170).

4 The bundle "a total of of patients" was extracted from fragments such as "a total of 128 of 212 patients".

5 The bundle "as compared with with placebo" was extracted from fragments such as "as compared with 44% with placebo".

References

Abdollahpour, Zeinab, and Javad Gholami. 2018. Building Blocks of Medical Abstracts: Frequency, Functions and Structures of Lexical Bundles. *The Asian ESP Journal* 14: 82–110. Available online: https://www.asian-esp-journal.com/volume-14-issue-1-june-2018/ (accessed on 18 August 2024).

Anderson, Kenneth, and Joan Maclean. 1997. A Genre Analysis Study of 80 Medical Abstracts. *Edinburgh Working Papers in Applied Linguistics* 8: 1–23.

Anthony, Laurence. 2023. AntConc (Version 4.2.4) [Computer Software]; Tokyo: Waseda University. Available online: https://www.laurenceanthony.net/software (accessed on 11 April 2024).
Anthony, Lawrence. 2024. AntWordProfiler. Available online: https://www.laurenceanthony.net/software/antwordprofiler/ (accessed on 1 August 2024).
Asano, Motoko, and Miho Fujieda. 2024. A Vocabulary Study for Enhancing Learners' Experiences: English-Language Medical Research Abstracts. *English Corpus Studies* 31: 1–22.
Asano, Motoko, Megumi Nakano, Yoshinori Miyazaki, and Miho Fujieda. 2022a. Introducing a Bilingual Corpus Database System of Medical Abstracts for Exploring Academic Connotations of Words: A Case Study of First-Year Medical Students. *Journal of Medical English Education* 21: 18–26.
Asano, Motoko, Megumi Nakano, Yoshinori Miyazaki, Tomoko Wakasa, and Miho Fujieda. 2022b. Use of Authentic Translation in Helping Students Decipher English-Language Randomised Control Trial Abstracts. *Journal of Medical English Education* 21: 87–93.
Ballance, Oliver, and Averil Coxhead. 2022. What can Corpora Tell us About English for Academic Purposes? In *The Routledge Handbook of Corpus Linguistics*, 2nd ed. Edited by Anne O'Keeffe and Michael J. McCarthy. London: Routledge, pp. 405–15.
Belcher, Diane D. 2016. English for Professional Academic Purposes. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Philip Shaw. Oxon: Routledge, pp. 502–14.
Bestgen, Yves. 2020. Comparing Lexical Bundles across Corpora of Different Sizes: The Zipfian Problem. *Journal of Quantitative Linguistics* 27: 272–90. [CrossRef]
Bhatia, Vijay K. 2019. Genre as Interdiscursive Performance in English for Professional Communication. In *Specialised English*. Edited by Ken Hyland and Lillian L. C. Wong. Oxon: Routledge, pp. 36–49.
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Essex: Longman.
Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. If You Look at . . .: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25: 371–405. [CrossRef]
Boers, Frank. 2020. Factors affecting the learning of multiword items. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. Oxon: Routledge, pp. 143–57.
Brooks, Gavin, Jon Clenton, and Simon Fraser. 2023. Exploring the Importance of Vocabulary for English as an Additional Language Learners' Reading Comprehension. In *EAL Research for the Classroom*. Edited by Gavin Brooks, Jon Clenton and Simon Fraser. Oxon: Routledge, pp. 35–58.
Browne, Charles. 2013. The New General Service List: Celebrating 60 Years of Vocabulary Learning. *The Language Teacher* 37: 13–16.
Browne, Charles. 2021. The NGSL Project: Building Wordlists and Resources to help EFL Learners (and Teachers) to Succeed. Paper Presented at JALTCALL 2020 Conference, Online, June 6–7; pp. 1–18. [CrossRef]
Browne, Charles, Brent Culligan, and Joseph Phillips. 2013. New Academic Word List. Available online: https://www.newgeneralservicelist.com/new-general-service-list-1 (accessed on 1 August 2024).
Casal, J. Elliott, and Matt Kessler. 2020. Form and Rhetorical Function of Phrase-frames in Promotional Writing: A Corpus- and Genre-based Analysis. *System* 95: 102370. [CrossRef]
Chen, Yu-Hua, and Paul Baker. 2010. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology* 14: 30–49.
Cook, Guy. 1998. The Uses of Reality: A Reply to Ronald Carter. *ELT Journal* 52: 57–63. [CrossRef]
Cook, Guy. 2010. *Translation in Language Teaching: An Argument for Reassessment*. Oxford: Oxford University Press.
Cortes, Viviana. 2004. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes* 23: 397–423. [CrossRef]
Cortes, Viviana. 2013. The Purpose of This Study is to: Connecting Lexical Bundles and Moves in Research Article Introductions. *English for Specific Purposes* 12: 33–43. [CrossRef]
Cortes, Viviana. 2015. Situating Lexical Bundles in the Formulaic Language Spectrum: Origins and Functional Analysis Development. In *Corpus-Based Research in Applied Linguistics: Studies in Honor of Doug Biber*. Edited by Viviana Cortes and Eniko Csomay. Amsterdam: John Benjamins, pp. 197–218.
Cortes, Viviana. 2024. Lexical Bundles and Phrase Frames. In *Conducting Genre-Based Research in Applied Linguistics*. Edited by Matt Kessler and Charlene Polio. New York: Routledge, pp. 105–26.
Coxhead, Averil. 2000. A New Academic Word List. *TESOL Quarterly* 34: 213–38. [CrossRef]
Coxhead, Averil. 2016. Acquiring Academic and Disciplinary Vocabulary. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Paul Shaw. Oxon: Routledge, pp. 177–90.
Coxhead, Averil. 2020. Academic Vocabulary. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. Oxon: Routledge, pp. 97–110.
Coxhead, Averil, and David Hirsh. 2007. A Pilot Science-specific Word List. *Revue Française de Linguistique Appliquée* 2: 65–78.
Coxhead, Averil, and Thi Ngoc Yen Dang. 2019. Vocabulary in University Tutorials and Laboratories. In *Specialised English*. Edited by Ken Hyland and Lillian L. C. Wong. Oxon: Routledge, pp. 120–34.
Culligan, Brent A. 2019. Evaluating Corpora with Word Lists and Word Difficulty. *Vocabulary Learning and Instruction* 8: 29–38. [CrossRef]
Dang, Thi Ngoc Yen. 2020. The Potential for Learning Specialized Vocabulary of University Lectures and Seminars through Watching Discipline-related TV Programs. *TESOL Quarterly* 54: 436–59. [CrossRef]

Dang, Thi Ngoc Yen, and Stuart Webb. 2016. Evaluating Lists of High-frequency Words. *International Journal of Applied Linguistics* 167: 132–58. [CrossRef]

Dang, Thi Ngoc Yen, Averil Coxhead, and Stuart Webb. 2017. The Academic Spoken Word List. *Language Learning* 67: 959–97. [CrossRef]

Durrant, Philip. 2017. Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories. *Applied Linguistics* 38: 165–93. [CrossRef]

EQUATOR Network. 2024. Reporting Guidelines. Available online: https://www.equator-network.org/ (accessed on 30 April 2024).

Flowerdew, John. 2000. Discourse community, Legitimate Peripheral Participation, and the Nonnative-English Speaking Scholar. *TESOL Quarterly* 34: 127–50. [CrossRef]

Flowerdew, Lynn. 2012. *Corpora and Language Education*. Hampshire: Palgrave MacMillan.

Fraser, Simon. 2007. Providing ESP Learners with the Vocabulary They Need: Corpora and the Creation of Specialized Word Lists. *Hiroshima Studies in Language and Language Education* 10: 127–43. [CrossRef]

Gardner, Dee, and Mark Davies. 2014. A New Academic Vocabulary List. *Applied Linguistics* 35: 305–27. [CrossRef]

Gilmore, Alexander, and Neil Millar. 2018. The Language of Civil Engineering Research Articles: A Corpus-based Approach. *English for Specific Purposes* 51: 1–17. [CrossRef]

Gilner, Leah. 2011. A Primer on the General Service List. *Reading in a Foreign Language* 23: 65–83.

Golparvar, Seyyed Ehsan, and Elyas Barabadi. 2020. Key Phrase Frames in the Discussion Section of Research Articles of Higher Education. *Lingua* 236: 102804. [CrossRef]

Graetz, Naomi. 1982. Teaching EFL Students to Extract Structural Information from Abstracts. Paper Presented at the International Symposium on Language for Special Purposes, Eindhoven, The Netherlands, August 2–4.

Gray, Bethany, Elena Cotos, and Jordan Smith. 2020. Combining Rhetorical Move Analysis with Multi-dimensional Analysis. In *Advances in Corpus-based Research on Academic Writing*. Edited by Ute Römer, Viviana Cortes and Eric Friginal. Amsterdam: John Benjamins, pp. 137–68.

Green, Clarence, and James Lambert. 2018. Advancing Disciplinary Literacy through English for Academic Purposes: Discipline-specific Wordlists, Collocations and Word Families for Eight Secondary Subjects. *English for Specific Purposes* 35: 105–15. [CrossRef]

Handford, Michael. 2010. What Can a Corpus Tell us About Specialist Genres? In *Routledge Handbook of Corpus Linguistics*. Edited by Anne O'Keeffe and Michael McCarthy. Oxon: Routledge, pp. 255–69.

Hanidar, Sharifah. 2016. Rhetorical Patterns, Verb Tense, and Voice in Cross Disciplinary Research Article Abstract. *Humaniora* 28: 12–27. [CrossRef]

Hartley, James. 1993. Recalling Structured Text: Does What Goes in Determine What Comes Out? *British Journal of Educational Tehnology* 24: 359–64. [CrossRef]

Hartley, James. 1999. Applying Ergonomics to Applied Ergonomics: Using Structured Abstracts. *Applied Ergonomics* 30: 535–41. [CrossRef]

Hartley, James. 2003. Improving the Clarity of Journal Abstracts in Psychology: The Case for Structure. *Science Communication* 24: 366–79. [CrossRef]

Huckin, Thomas N., and Leslie A. Olsen. 1983. *English for Science and Technology: A Handbook for Nonnative Speakers*. New York: McGraw-Hill.

Hyland, Ken. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Hyland, Ken. 2008a. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* 27: 4–21. [CrossRef]

Hyland, Ken. 2008b. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics* 18: 41–62. [CrossRef]

Hyland, Ken. 2012. Disciplinary Specificity: Discourse, Context, and ESP. In *New Directions in English for Specific Purposes Research*. Edited by Diane Belcher, Ann M. Johns and Brian Paltridge. Ann Arbor: The University of Michigan Press, pp. 6–24.

Hyland, Ken. 2016. General and Specific EAP. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Philip Shaw. Oxon: Routledge, pp. 17–29.

Hyland, Ken, and Feng (Kevin) Jiang. 2019. *Academic Discourse and Global Publishing: Disciplinary Persuasion in Changing Times*. London: Routledge.

Hyon, Sunny. 2018. *Introducing Genre and English for Specific Purposes*. London: Routledge.

ICMJE. 2024. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Available online: https://www.icmje.org/icmje-recommendations.pdf (accessed on 30 April 2024).

Imao, Yasuhiro. 2024. CasualConc (Version 3.0.8). [Computer Software]. Available online: https://sites.google.com/site/casualconcj/ (accessed on 9 July 2024).

Jego, Eric Hajime. 2012. Corpus Analysis Demonstrates That Scientific Writing Uses the Structure "<Disease> was Diagnosed. . ." More Than "<Person> was Diagnosed. . .". *The Journal of Medical English Education* 11: 50–58.

Kanoksilapatham, Budsaba. 2005. Rhetorical Structure of Biochemistry Research Articles. *English for Specific Purposes* 24: 269–92. [CrossRef]

Kanoksilapatham, Budsaba. 2009. Generic Structure of Research Article Abstracts in Sciences. *Journal of English Studies* 4: 95–111. Available online: https://so04.tci-thaijo.org/index.php/jsel/article/view/22067 (accessed on 18 August 2024).

Lake, William Michael, and Viviana Cortes. 2020. Lexical Bundles as Reflections of Disciplinary Norms in Spanish and English Literary Criticism, History, and Psychology Research. In *Advances in Corpus-Based Research on Academic Writing*. Edited by Ute Römer, Viviana Cortes and Eric Friginal. Amsterdam: John Benjamins, pp. 183–203.

Li, Liang, Margaret Franken, and Shaoqun Wu. 2020. Bundle-driven Move Analysis: Sentence Initial Lexical Bundles in PhD Abstracts. *English for Specific Purposes* 60: 85–97. [CrossRef]

Liu, Chen-Yu, and Hao-Jan Howard Chen. 2022. A Phraseological Exploration of University Lectures through Phrase Frames. *Journal of English for Academic Purposes* 58: 101135. [CrossRef]

Luo, Na, and Ken Hyland. 2019. "I Won't Publish in Chinese Now": Publishing, Translation and the Non-English Speaking Academic. *Journal of English for Academic Purposes* 39: 37–47. [CrossRef]

MacDonald, Ewen. 2019. An Analysis of Vocabulary Level in Reading Passages of the National Center Test. *Shiken* 23: 19–27.

Maswana, Sayoko, Toshiyuki Kanamaru, and Akira Tajino. 2015. Move Analysis of Research Articles Across Five Engineering Fields: What They Share and What They Do Not. *Ampersand* 2: 1–11. [CrossRef]

Millar, Neil, Brian Budgell, and Keith Fuller. 2012. 'Use the Active Voice Whenever Possible': The Impact of Style Guidelines in Medical Journals. *Applied Linguistics* 34: 393–414. [CrossRef]

Millar, Neil, Françoise Salager-Meyer, and Brian Budgell. 2019. It Is Important to Reinforce the Importance of. . .": 'Hype' in Reports of Randomized Controlled Trials. *English for Specific Purposes* 54: 139–51. [CrossRef]

Mizumoto, Atsushi. 2015. Corpus-based Analysis of Lexical Bundles: Its Potential Applications in English Language Teaching. *Journal of the Japan Society for Speech Sciences* 16: 30–34.

Mizumoto, Atsushi, Geoffrey G. Pinchbeck, and Stuart McLean. 2021. Comparisons of Word Lists on New Word Level Checker. *Vocabulary Learning and Instruction* 10: 30–41. [CrossRef]

Mizumoto, Atsushi, Sawako Hamatani, and Yasuhiro Imao. 2017. Applying the Bundle–Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles. *Language Learning* 67: 885–921. [CrossRef]

Moreno, Ana I., and John M. Swales. 2018. Strengthening Move Analysis Methodology towards Bridging the Function-Form Gap. *English for Specific Purposes* 50: 40–63. [CrossRef]

Morley, John. 2023. Academic Phrasebank. Available online: https://phrasebankresearch.net/ (accessed on 4 July 2024).

Nam, Sejin, Senator Jeong, Sang-Kyun Kim, Hong-Gee Kim, Victoria Ngo, and Nansu Zong. 2016. Structuralizing Biomedical Abstracts with Discriminative Linguistic Features. *Computers in Biology and Medicine* 79: 276–85. [CrossRef]

Nankodo. 2024. The New England Journal of Medicine Nihon Kokunai Ban [The Japanese Edition]. Available online: https://nejm.jp (accessed on 25 April 2024).

Nation, Ian Stephen Paul. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, Ian Stephen Paul. 2005. Teaching and learning vocabulary. In *Handbook of Research in Second Language Teaching and Learning*. Edited by Eli Hinkel. Mahwah: Lawrence Erlbaum Associates, Inc., pp. 581–95.

Nation, Ian Stephen Paul. 2022. *Learning Vocabulary in Another Language*, 3rd ed. Cambridge: Cambridge University Press.

Noto, Hitoshi. 2015. The New England Journal of Medicine (NEJM) Saito no Katsuyo [The Effective Use of the New England Journal of Medicine Websites]. *Hosupitaru Raiburarian [Hospital Librarian]* 40: 159–61. (In Japanese).

Nwogu, Kevin N. 1997. The Medical Research Paper: Structure and Functions. *English for Specific Purposes* 16: 119–38. [CrossRef]

Ogawa, Ryuichi. 2014. Daigaku yakugakubu ni okeru kyoiku [Education in university pharmacy schools]. *The Japanese Journal of Applied Therapeutics* 6: 41–46. [CrossRef]

Omidian, Taha, Hesamoddin Shahriari, and Anna Siyanova-Chanturia. 2018. A Cross-disciplinary Investigation of Multi-word Expressions in the Moves of Research Article Abstracts. *Journal of English for Academic Purposes* 36: 1–14. [CrossRef]

Pho, Phuong Dzung. 2008. Research Article Abstracts in Applied Linguistics and Educational Technology: A Study of Linguistic Realizations of Rhetorical Structure and Authorial Stance. *Discourse Studies* 10: 231–35. [CrossRef]

Qi, Hui, and Fan Pan. 2020. Lexical Bundle Variation Across Moves in Abstracts of Medical Research Articles. *Southern African Linguistics and Applied Language Studies* 38: 109–28. [CrossRef]

Quero, Betsy, and Averil Coxhead. 2018. Using a Corpus-based Approach to Select Medical Vocabulary for an ESP Course: The Case for High-frequency Vocabulary. In *Key Issues in English for Specific Purposes in Higher Education*. Edited by Yasemin Kirkgöz and Kenan Dikilitaş. Berlin and Heidelberg: Springer, pp. 51–75.

Rees, Geraint. 2022. Using Corpora to Write Dictionaries. In *The Routledge Handbook of Corpus Linguistics*, 2nd ed. Edited by Anne O'Keeffe and Michael J. McCarthy. London: Routledge, pp. 387–404.

Reppen, Randi, and Shannon B. Olson. 2020. Lexical Bundles Across Disciplines. In *Advances in Corpus-Based Research on Academic Writing*. Edited by Ute Römer, Viviana Cortes and Eric Friginal. Amsterdam: John Benjamins, pp. 169–82.

Robinson, Marin S., Fredricka L. Stoller, Molly S. Costanza-Robinson, and James K. Jones. 2008. *Write Like a Chemist*. Oxford: Oxford University Press.

Salager-Meyer, Françoise. 1991. Medical English Abstracts: How Well Are They Structured? *Journal of the American Society for Information Science* 42: 528–31. [CrossRef]

Salager-Meyer, Françoise. 1992. A Text-type and Move Analysis Study of Verb Tense and Modality Distribution in Medical English Abstracts. *English for Specific Purposes* 11: 93–113. [CrossRef]

Salager-Meyer, Françoise, María Ángeles Alcaraz Ariza, and Beverly A. Lewin. 2014. Abstract Quality in Complementary and Alternative Medicine Papers: A Structural and Cross-Generic Analysis. In *Abstracts in Academic Discourse*. Edited by Marina Bondi and Rosa Lorés Sanz. Bern: Peter Lang, pp. 221–40.

Samraj, Betty. 2016. Research Articles. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Philip Shaw. Oxon: Routledge, pp. 403–15.

Sarica, Serhad, and Jianxi Luo. 2021. Stopwords in Technical Language Processing. *PLoS ONE* 16: e0254937. [CrossRef]

Shimizu, Maki. 2019. Japanese medical students' reading of English academic papers and an evaluation of their ability to put grammatical knowledge to practical use. *Journal of Medical English Education* 18: 82–86.

Simpson, Alan. 2022. Medical Students' English Needs and Curricular Developments. *Journal of Medical English Education* 21: 94–100.

Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31: 487–512. [CrossRef]

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: The Oxford University Press.

Stoller, Fredrica L., and Marin S. Robinson. 2013. Chemistry Journal Articles: An Interdisciplinary Approach to Move Analysis with Pedagogical Aims. *English for Specific Purposes* 32: 45–57. [CrossRef]

Storch, Neomy, Janne Morton, and Celia Thomson. 2016. EAP Pedagogy in Undergraduate Contexts. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Philip Shaw. Oxon: Routledge, pp. 477–88.

Stosic, Dragana. 2022. A Linguistic Construction of a Research Gap in Reports of Randomised Controlled Trials: A Systemic Functional Perspective. *Journal of English for Academic Purposes* 59: 101158. [CrossRef]

Stubbs, Michael, and Isabel Barth. 2003. Using Recurrent Phrases as Text-Type Discriminators: A Quantitative Method and Some Findings. *Functions of Languages* 10: 65–108. [CrossRef]

Swales, John. 1981. *Aspects of Article Introductions*. Birmingham: The Language Studies Unit, The University of Aston in Birmingham.

Swales, John M. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.

Swales, John M. 2004. *Research Genres*. Cambridge: Cambridge University Press.

Swales, John M. 2017. Standardisation and its Discontents. In *Publishing Research in English as an Additional Language: Practices, Pathways, and Potentials*. Edited by Margaret Cargill and Sally Burgess. Adelaide: University of Adelaide Press, pp. 239–54.

Tang, Chiachieh, and Yeu-Ting Liu. 2019. Construction and Proposal of an Academic Word List for Medical Professionals at Different Developmental Stages. *Journal of Medical English Education* 18: 13–20.

Tankó, Gyula. 2017. Literary Research Article Abstracts: An Analysis of Rhetorical Moves and Their Linguistic Realizations. *Journal of English for Academic Purposes* 27: 42–55. [CrossRef]

Tardy, Christine M. 2009. *Building Genre Knowledge*. Anderson: Parlor Press.

Tardy, Christine M., and Soomin Jwa. 2016. Composition Studies and EAP. In *The Routledge Handbook of English for Academic Purposes*. Edited by Ken Hyland and Philip Shaw. Oxon: Routledge, pp. 56–68.

The New England Journal of Medicine. 2024. Available online: https://nejm.org/ (accessed on 18 August 2024).

Thorndike, Edward Lee. 1921. *The Teacher's Word Book*. New York: Teachers College, Columbia University.

Thorndike, Edward Lee, and Irving Lorge. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University. Available online: https://books.google.com/books?id=FeJZAAAAMAAJ&q=The+Teacher%E2%80%99s+Word+Book+of+30,000+Words&dq=The+Teacher%E2%80%99s+Word+Book+of+30,000+Words&hl=ja&newbks=1&newbks_redir=1&printsec=frontcover&sa=X&ved=2ahUKEwjE9ZTW8IqIAxWkjK8BHTfBAI4Q6AF6BAgJEAI (accessed on 18 August 2024).

Tribble, Christopher. 2017. ELFA vs. Genre: A New Paradigm War in EAP Writing Instruction? *Journal of English for Academic Purposes* 25: 30–44. [CrossRef]

Viana, Vander, and Aisling O'Boyle. 2022. *Corpus Linguistics for English for Academic Purposes*. Oxon: Routledge.

Wang, Jing, Shao-lan Liang, and Guan-chun Ge. 2008. Establishment of a Medical Academic Word List. *English for Specific Purposes* 27: 442–58. [CrossRef]

West, Michael. 1953. *A General Service List of English Words*. London: Longman.

Widdowson, Henry G. 1991. The Description and Prescription of Language. In *Linguistics and Language Pedagogy: The State of the Art*. Edited by James Alatis. Washington, DC: Georgetown University Press, pp. 11–24.

Zhang, Chunfang, and Xueli Liu. 2011. Review of James Hartley's Research on Structured Abstracts. *Journal of Information Science* 37: 570–76. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Size Matters: Vocabulary Knowledge as Advantage in Partner Selection

Michael Daller [1,*] and Zehra Ongun [2]

[1] Department of English Language and Applied Linguistics, University of Reading, Reading RG6 6UR, UK
[2] School of Culture and Communication/English Language, TESOL, Applied Linguistics, University of Swansea, Swansea SA2 8PP, UK; zehra82karaman@hotmail.com
* Correspondence: m.daller@reading.ac.uk

Abstract: Partner selection can be studied from different disciplines, such as psychology, sociology, and economics. However, linguistic perspectives have been neglected. That is why we need an interdisciplinary approach that includes language. The present article investigates how important the vocabulary size of a potential partner is for marital choice. Our theoretical framework is mainly that of biological markets which are still being widely used. This framework assumes that human decisions are made on a rational basis, e.g., about the characteristics that a potential partner brings into a marriage such as economic assets (wealth, education), psychological traits (intelligence, kindness, fairness), or signs that show physical and mental health. Partner selection takes place on a biological market where assets are displayed and are part of the negotiation for the best partner. We argue that vocabulary knowledge is such an asset, which is acquired through lengthy and costly education and distinguishes potential partners (or their parents) who can afford the accumulation of this form of human capital. Markets are not fully transparent and our knowledge about a potential partner might be incomplete or even distorted through false information or even cheating as one can clearly see from advertisements in online dating. However, we cannot pretend, at least not over a longer period of time, to know words that are not at our disposal. This present study is based on data from 83 couples after more than 15 years of marriage. Their vocabulary scores correlate highly and it is possible that this correlation is the result of accommodation through marriage. However, through partialling out statistically the years of marriage we conclude that the vocabulary size of each partner was an important factor already right at the beginning of their relationship. Those with higher human capital in vocabulary attract similar partners, and this holds for males and females as well as vice versa. Our participants are all Turkish–English sequential bilinguals and the question is whether it is vocabulary knowledge in the first or the second language that plays a crucial role in partner selection. Our results show that both languages are important. We argue that it is not knowledge of words at the surface level but that it is knowledge of conceptual concepts underlying both languages that serve as a display of human capital on the biological market of partner selection.

Keywords: vocabulary size; infrequent vocabulary knowledge; partner selection

1. Introduction

The present article investigates how important the vocabulary size of a potential partner is for marital choice. Our theoretical framework is mainly that of biological markets (Becker 1976, et passim), which are still being widely used (Becker et al. 2023). This framework assumes that human decisions are made on a rational basis, e.g., about the characteristics that a potential partner brings into a marriage such as economic assets (wealth, education), psychological traits (intelligence, kindness, fairness), or signs that show physical and mental health (see Becker 2007). To make decisions about a partner we need to know which assets or traits are brought into a potential marriage. However, biological markets are not fully transparent and our knowledge about a potential partner

might be incomplete or even distorted through false information or even cheating, as one can clearly see from advertisements in online dating (see, for example, Ranzine and Lutz 2017; Toma and Hancock 2010; Nascimento et al. 2024).

In biology, the concept of an honest signal is used to show how distortion of information is avoided. A classic example is the peacock's tail. It shows the fitness of the male and is used as a biological market of partner selection. It is honest, as the peacock cannot fake it and show feathers that it does not have. We argue that the appropriate use of infrequent vocabulary is a similar honest signal and one cannot lie about word knowledge that one does not possess. The acquisition of infrequent vocabulary is time-consuming and costly through education and its correct display cannot be fabricated, at least not over a longer period of time. Vocabulary knowledge as human capital (see also Bourdieu 2018) is a proxy for the potential assets of the partner.

Our sample consists of 83 sequential Turkish couples who were raised in Turkey but moved to the UK and use English on a day-to-day basis. These couples were used in an earlier study (Daller and Ongun 2018) about the relationship between the vocabulary sizes of parents and their children. We had only access to mixed-sex partners with children. The reason for selecting bilingual parents and children in the earlier study was to obtain further insight into the cognitive advantages or disadvantages of vocabulary knowledge in two languages. In this present study, the use of bilingual participants can give us insights into whether the market value of a potential partner is reflected in both languages or only in one (for details, see Section Hypotheses).

We use a yes–no format test (X-lex, Meara and Milton 2003), that includes non-existing words to control for dishonest signals by candidates who claim to know these non-words. We administered tests in both languages of the participants to be able to draw a fine-grained picture of their vocabulary knowledge.

If potential partners use their vocabulary in mutual selection, correlations between their test scores at least in one language of our bilingual sample can be expected. This could be explained in two different ways: (a) the couples made their decision to select each other at the beginning of their relationship based on the vocabulary sizes displayed or (b) they made their decision based on other criteria and adapted to each other's vocabulary size over the years of marriage. A decision about these two explanations can be made by partialling out statistically the length of marriage[1]. This can give us further insights in how quickly decisions about partner selection are made.

2. Literature Review

2.1. Theories of Partner Selection

According to Becker (1976, 2007, 2009) and Becker et al. (2023) all human behaviour can be explained with economic models, which is a view that is still widely held (see Heckman et al. 2018; Barclay 2013, 2016) but has also been criticised (see the end of this section). It has been argued in favour of Becker's framework that "the constraints of life force every organism to behave economically" (Noë et al. 2001, p. ix.). This means that partner choice can also be explained in economic terms, where, according to Noë and Hammerstein (1995, p. 336), decisions on cooperation and sexual selection are made on a biological market.

In the present context, we focus entirely on human marital partner selection, but we are aware that the concept of biological markets is used in a large variety of other contexts, i.e., partner choice in economic games (Eisenbruch and Roney 2017), in the choice of campmates in hunter–gatherers in Tanzania (Smith and Apicella 2020), or in a variety of partner selection in mammals (Dunbar 2001), to name a few.

Becker suggests that two principles are the basis of marital choice: (a) it is a voluntary decision either made by the partners or their parents based on the assumption that marriage would increase the available assets of the partners, and (b) that "since many men and women compete . . . a market in marriages can be presumed". (see Becker 1976, p. 206). Each person tries to find the best mate on that market. Becker (1976, p. 217) calls this

assertive mating where a set of traits such as intelligence, education, height, and tendency to nurture (see Becker 2009) are displayed.

According to Becker, the main driving force behind marriage is that it "maximizes total household commodity" (Becker 1976, p. 217), which also includes "commodities" such as companionship and health. One has to bear in mind that Becker's view on commodities is general in nature and goes beyond resources that are traditionally considered in economics. In a more traditional view, "love" is not seen as an economic commodity, but it can be argued that "caring" has an economic aspect, which can be easily seen when one considers the cost of a caring home. Becker argues in economic terms but these commodities cannot always be expressed directly in a financial way. In economics, the success of a business can be qualitatively measured in the turnover of the company, which is a result of the trust of the customers or partners in this business. Trust in a partnership is also an asset for the success of a relationship but a monetary quantification would be more complex, although it would be measurable in principle.

Becker's theoretical framework has been criticised as neo-liberal, and alternative explanations of human behaviour are discussed in economic science (see Lavoie 1992; Deane 2020). Becker's work has also been criticised because it applies explanations in areas that fall in the realm of other disciplines, such as sociology (see Coleman 1993). It has been argued that human behaviour is not entirely based on economic determinants only but also on the influence of culture (see Taylor 1994). The concept of a biological market is linked to the rational choice theory, which assumes that humans can make rational decisions based on the evaluation of the costs and benefits of their choices (see Scott 2000). Within economics, this theory is not uncontroversial. Hodgeson (2012) argues that this theory is too abstract and that we "have to consider the real social and psychological determinants of human behaviour" (Hodgeson 2012, p. 94). Hodgeson (2003) describes in detail how economic decisions, i.e., to buy something, can be manipulated and influenced by persuasion. We agree that real people are more than a homo economicus and that many aspects play a role in their decision making, but it is very unlikely that humans make their decisions at random. It is just that we do not understand every determinant of the decision-making process. According to the efficient market theory (for a critical discussion, see Delcey 2019), in an ideal market, all information is available to every participant and the market is totally transparent, which is of course only an ideal without undue manipulation or insider knowledge, which might be a criminal offence (see Ahern 2017).

It is beyond the scope of the present article to discuss the economic theories mentioned above in detail, but any behavioural science, including applied linguistics, can learn from other disciplines. An important factor in human decision making is trustworthy knowledge about the potential partner and its contribution to a possible relationship. Potential partners will try to avoid manipulation and cheating and this means that the biological market, as every functioning market, needs to be as transparent as possible.

2.2. The Process of Partner Selection on a Biological Market

Noë and Hammerstein (1995) compare potential partners with traders that try to outbid their competitors in a market. In a similar vein, and in line with Becker's theoretical framework, Barclay (2013) argues that when potential partners can choose with whom they can interact, a biological market is created and individuals will try to outbid their rivals (Barclay 2013, p. 164). The question is how this "outbidding" works in practice. Barclay (2016, p. 33) argues that cooperation is crucial to social success and that we avoid "bad partners" and try to associate with "better ones". This creates a biological market where people try to appear to be able to confer more benefits than others to their partner "causing an arms race over any behaviour that signals these traits" (Barclay 2016, p. 33). In this "arms race" misleading and cheating is a possibility, as can be seen in false information in online dating, e.g., about one's own age, education, or wealth (see Ranzine and Lutz 2017; Toma and Hancock 2010). The discussion about the psychological aspects of partner selection is ongoing (Devenport et al. 2023) but the concept of biological markets is certainly an

important aspect of human decision making in this field. The question is how a transparent market can be created where the possibilities of cheating are avoided.

2.3. Transparency through Honest Signals

The concept of "honest" or "costly" signaling can give us some insights into mate selection in animals (and humans) (see Pentland 2010). According to Zahavi and Zahavi (1999), a peacock's tail can act as a signal about the health and genetic fitness of the animal. This is because "Unhealthy, unfit peacocks can't afford big, bright tails" (Miller 2000, p. 64), as they are very "costly" and lots of energy and food are necessary to develop such a tail that is not necessary for the survival of the individual bird.

A peacock cannot make a conscious decision about the number of feathers or their form. It cannot deceive the peahen who makes her mating decision based on this honest signal. Human language traits such as fluency and automaticity can be linked to the theory of honest signaling. The comparison with a peacock is of course a simplification, but, in our view, it remains a useful metaphor. Adults can still increase their vocabulary knowledge and fluency over time, whereas a peacock probably cannot increase the number of its feathers. However, what is relevant here is the vocabulary knowledge at the time of partner selection where one cannot deceive and speak more fluently than one's own abilities allow. General intelligence can also be linked to honest signals in humans (Luxen and Buunk 2006). One cannot convincingly appear more intelligent than one is, at least not over a longer period.

Vocabulary knowledge is related to crystallised intelligence (see, for example, Stamenković et al. 2019). It is about knowing things as opposed to fluid intelligence, which is needed for solving tasks in new situations, e.g., pattern recognition. Crystallised intelligence distinguishes the speakers from those (or their parents) who did not have the resources to develop this human capital (see Bourdieu 2018) through (costly) schooling and education (see also Brown 2010; Milton and Treffers-Daller 2013).

Language can be used to cheat and convey false information. In fact, lying is in many cases only possible with language (for a discussion about lying, see García-Carpintero 2023). Cheating with vocabulary **knowledge**, however, is difficult, as it is not possible to pretend to know unknown words convincingly over a longer period. Vocabulary display is therefore comparable to a peacock, which cannot "cheat" about the richness of its feathered tail at will, nor can humans cheat about their knowledge of infrequent words. This is why vocabulary tests based on the yes–no format include non-existing words. If a candidate claims that they know a non-word, they are probably cheating, and their test scores are adjusted accordingly in this format, e.g., in the X-lex test format that we use in this present study (see Meara and Milton 2003).

One question is why we develop large vocabularies that are not strictly necessary in everyday life. According to Miller (2000), the development of large vocabularies is an evolutionary process. "Over many generations, those with quicker wits and more generous spirits may have attracted more sexual partners" (Miller 2000, p. 10). To be wittier, one needs to have a greater variety in one's vocabulary. According to Miller (2000, p. 133), larger vocabularies have led to higher mating success in our ancestors, and "they would have benefited by evolving larger vocabularies, just as peacocks evolved larger tails" (Miller 2000, p. 373). According to Miller, both female and male partners are involved and "verbal courtship is mutual" (Miller 2000, p. 383). In summary, the display of real vocabulary knowledge is an honest signal about one's own human capital and it is important in mating success on the biological market of partner selection.

2.4. Empirical Studies in Partner Selection

In the following section, we will discuss previous studies that investigate the process of partner selection on an empirical basis. There are two principal empirical approaches to investigate mate preferences: (a) large-scale studies based on databases or questionnaires and (b) experimental designs. Haandrikman and van Wissen (2012) carried out a study

with 326,000 people who, according to the Dutch population register, had recently started living together. In their study, they show that several factors are important for partner selection, including similar education, income level, dialect or language, and the same social class. This is in line with other large-scale studies on databases (i.e., Kalmijn 1991). In fact, dissimilarities between partners increase the risk of divorce (Dupont et al. 2020). Such (dis)similarities can be determined through questionnaires about partner preferences, as in Buss et al. (1990), who carried out a study in 37 different cultures with 9474 participants in total. Their aim was to identify psychological traits that are identical across cultures for partner choice. They asked the participants to rate personal characteristics according to their importance for partner selection (one 18-item and one 72-item questionnaire). With several hierarchical regressions, they identified characteristics that are more culture-specific and characteristics that are important across cultures. An example for culture-specific characteristics is religious preferences. There are also aspects that are more relevant within certain clusters of cultures, such as chastity with no previous experience in sexual intercourse. China or India, for example, place great value on this characteristic, whereas it was seen as unimportant, for example, in (then) West Germany in this study. More important for a general view are those characteristics that are relevant across cultures. Income and education seem to be more important in males, and physical attractiveness in females, which is in line with general stereotypes about mate choices and mate retention strategies (see also Nebl et al. 2021). But across cultures, males and females place the first four characteristics in the same order: kind and understanding, intelligent, exciting personality, and health (see Buss et al. 1990, p. 18).

Fisman et al. (2006) investigate partner choice in a speed dating experiment at a Western university. The 392 undergraduate and postgraduate students who took part in the study provided judgements on a series of characteristics of their preferred interlocutors. The authors come to the conclusion that males look more for physical characteristics in females, whereas females find males that are intelligent more attractive (Fisman et al. 2006, pp. 682–83). One other aspect that is important for both genders is ambition. This might, however, be a specific characteristic of a student sample. Pawlowski and Koziel (2002) analysed 2008 personal advertisements for partner choice. They found that for men, the best indicator for the number of responses to the advertisements is their educational level. Education is assumed to be related to intelligence (i.e., Deary et al. 2007; Hegelund et al. 2020) and socio–economic status. Personal advertisements on online dating apps were also investigated by Van der Zanden et al. (2020). They found that spelling mistakes on dating websites have a negative effect on the recipient, whereas the use of metaphoric and therefore varied language has a positive one.

Education and intelligence are clearly important in human mating selection (see also Geher et al. 2020). This then leads to the question of how a person can judge the intelligence of their interlocutor. Further insights into the ways in which individuals display their market value can be obtained from Rosenberg and Tunney (2008), who first primed 83 (33 males, 53 females) undergraduate students with a set of photos from the opposite sex (younger and older faces) and then asked their participants to describe their first romantic encounter at the university (not with the person in the photo; this was just a prime). Males used significantly more rare words when they were primed with a younger female face than with an older face. Females judged showed a similar picture but the difference was only a statistical trend. However, this seems to be an indication that vocabulary is important in mutual partner selection.

This is an indication that males display their vocabulary in romantic encounters (see Miller in the previous section). Rosenberg and Tunney explain their findings with the assumption that "vocabulary is a direct indicator of intelligence" and therefore an important criterion in partner selection (Rosenberg and Tunney 2008, p. 539).

The question is how fast decisions about these aspects can be made in a potential partner. Part of the design of speed dating (see Fisman et al. 2006) is the short time that the participants are given to make their decision, as each encounter lasts only four minutes

in this study. It is even possible that the decisions were made at the beginning of this four-minute period. Humans can obviously make decisions about their interlocutor in a short period of time. This is also relevant in contexts other than partner selection, such as job interviews. Frieder et al. (2016) analysed 691 applicants at a university career centre. The interviews were scheduled for 30 min, but 30.4% of the interviewers reported that they made the decision by the fifth minute in the interview, and 40.1% reported that they made the decision around the fifteenth minute. Likewise, there does not seem to be a relationship between job interview length, interview reliability, and validity (Thorsteinson 2018). To our knowledge, there are no studies that investigate the role of vocabulary in judgements on interlocutors in these contexts. Some evidence can be found in studies on teacher judgements of language learners. Meara and Babi (2001), for example, presented assessors of English with text written by native speakers and non-native speakers. The assessors saw the texts one word at a time in a cumulative manner. They were then asked to indicate at what point they felt confident to make a judgement on the texts. Typically, they said that they were confident whether a text was written by a native or non-native speaker after they read 30 words, and some of them needed even fewer words. However, Meara and Babi also admit that it is unclear how the assessors arrived at their decisions in such a short time (Meara and Babi 2001, p. 80). Although this study investigates written texts, some conclusions can be drawn for the evaluation of speech. According to Meara and Babi (2001, p. 82), about 30 written words are the equivalent of 10 s of speech, which means that it is possible to judge a speaking person after a very short period.

In our sample (see below), we have sequential bilinguals who use their two languages on an everyday basis. This makes the research design more complex, as vocabulary in the first language might indicate different traits than vocabulary in the second language, e.g., intelligence, which might be closer related to the first language and early cognitive development (Clark and Casillas 2015) versus economic success, which is hardly possible without profound knowledge of the language of the environment that is the second language of the participants. Vocabulary in the first language will differ from that in the second, according to the complementary principle (Grosjean [1982] 2001, 2015), as bilinguals use their languages in different domains and "balanced" bilinguals with equal vocabularies in both of their languages hardly exist (see Treffers-Daller and Silva-Corvalán 2016). However, there are indications that at least the receptive vocabularies of a bilingual are related and that there are some common psychological traits underlying vocabularies in both languages (see Cummins 1976, 1991). These common underlying traits are related to factors such as conceptual knowledge or crystallised intelligence, which is world knowledge based on experience and education (for a summary of the discussion on a common underlying proficiency see (Daller and Ongun 2018). The focus of this present paper is partner selection and not a detailed discussion of the complex, psychological relationship between the vocabularies of bilinguals. However, this present study gives us deeper insights into the role of bilingual vocabulary displays as indicators of the market value of a bilingual.

Overall, the literature overwhelmingly shows that intelligence, education, and socio–economic status are important factors on the biological market of partner selection across a range of different cultures. Decisions in a partner selection process might be short-termed, but even if there is more time available, it is difficult to evaluate fully which assets or commodities (in Becker's terms) are offered by a candidate as markets are normally not fully transparent, and there is the possibility of false information or even cheating. Market transparency can be sought by using honest signals that cannot be faked. The appropriate use of infrequent vocabulary is such an honest signal and the use of a varied vocabulary is an important factor on the biological market of partner selection.

3. Research Questions

3.1. Is the market value of males more important than those of females in partner selection or is partner selection a mutual process between both partners, as suggested by Miller (2000, p. 383)?[2]

3.2. In successive bilinguals, one language (usually called the first language) is the main means of education, and this language might be the main indicator of the market value of a partner. However, it can be assumed that there are some psychological traits underlying both languages, e.g., conceptual knowledge as an indicator of experience and education. Daller and Ongun (2018) discuss this research field in more detail and come to the conclusion that in bilinguals, vocabulary in both languages combined reflect common underlying traits such as education and conceptual knowledge. The question in this present study is whether the first language, the second language, or both are indicators of the market value of a potential partner.

3.3. As the literature shows, judgements about the characteristics of a possible interlocutor are made quite quickly, e.g., in job interviews, speed dating, or teacher judgements (Frieder et al. 2016; Meara and Babi 2001; Thorsteinson 2018). We, therefore, assume that in partner selection, judgements are also made quite quickly at the beginning of a relationship. To obtain deeper insight into this decision-making process, we need to know whether similarities in vocabulary sizes are the result of accommodation through marriage or if they have existed already at the time of partner selection as the basis for a possible judgement.

Hypotheses

H1: *Vocabulary size does not play a role in partner selection and therefore there will be no correlation between the vocabulary sizes of the couples in our sample.*

H1 alternative: *There will be a high correlation in the vocabulary sizes of the participants because vocabulary size reflects intelligence, education, and the market value of a person (see Geher et al. 2020), and mutual courtship is assumed (see Miller 2000, p. 383).*

H2a: *In the same couple, the male displays a larger vocabulary than the female partner. This assumption is based on some literature findings that indicate that males adopt their speech more with rare vocabulary when primed with younger female faces (Rosenberg and Tunney 2008).*

H2b: *In the same couple, the female displays a larger vocabulary than the male partner.*

H2 alternative: *In the same partnership, both sexes have similar vocabulary sizes because on a biological market, traits such as intelligence and education (see Becker 2009) are selected based on mutual courtship (see Miller 2000, p. 383).*

H3: *Only vocabulary size in one language correlates between the two partners.*

H3 alternative: *Vocabulary in both languages of the two partners correlates positively as there are common underlying psychological traits between the languages of a bilingual (see Cummins 1976, 1991; Daller and Ongun 2018).*

H4: *The length of marriage correlates with the vocabulary sizes of both partners.*

H4 alternative: *The length of marriage (partialled out statistically) does not correlate between the vocabulary sizes of the partners because the judgement on a potential partner is made quickly (see summary in Section 2.3) and any possible accommodation through marriage is no longer relevant for the initial partner selection.*

4. Methodology

4.1. Participants

The participants in this study are 83 Turkish–English bilingual married couples living in the UK. Originally, they were selected as parents for a larger study on the bilingual advantage in children (see Daller and Ongun 2018). In this earlier study, we studied the

vocabulary of 100 children of the 83 married couples, but now the focus is on the parents only. The length of the marriages ranges from 9 to 23 years with a mean of 15.46 years (*St.Dev.* 3.46). Their ages varied between 27 and 57 (M = 31, *St.Dev.* = 1.21). They are from the middle class (SES) as at least one partner of each couple has a higher education degree (college or university). All participants have Turkish as their first language. They were born in Turkey and had come to the UK after finishing their education in Turkey. They use English in their daily life, but in many cases, the home language is still Turkish.

4.2. Measures

The vocabulary size of the participants was measured with an X-lex test in English and in Turkish. The X-lex is a receptive vocabulary size test developed by Meara and Milton (2003), designed in a yes–no format. The test comprises 100 real words and 20 pseudo or non-words. From each frequency band from 1 k to 5 k, 20 real words are selected at random. The frequency list that the authors used is drawn from Nation's (1984) and Hindmarsh's (1980) frequency lists.

The pseudo- or non-words resemble real words but do not exist. These are included to reduce the effect of guesswork and overestimation. Examinees are required to tick the words that they know; each tick for a real word is counted, multiplied by 50, and added to the score, but each tick for a pseudo or non-word is multiplied by 250 and deducted from the total score. This means that if all real words but no pseudo-words are accepted the maximum score will be 5000. In the case of pure guessing, the score will tend to become zero. The Turkish X-lex test was developed by the second author of this study according to the existing X-lex format in English. Pseudo-words were included that are phonologically plausible in Turkish but do not exist. We used a random number allocator to determine the distribution of the pseudo-words in the test. In a pilot study, it turned out that the original Turkish test was too easy for native speakers and that there was a clear ceiling effect with almost 100% correct scores. We therefore included more infrequent words from frequency bands up to 15 k. Again, the maximum score was 5000.

4.3. Procedure

We decided to start always with the English test and two weeks later with the Turkish test without any specific reason for this order, keeping the test setting constant. The English and Turkish X-Lex tests were administered to the partners separately, and the other partner was not in the same room when the test was taken. The informants were provided with a clear oral instruction regarding the format of the test and the purpose of the task in English and Turkish to avoid priming effects. Each test took 10 min to complete.

5. Results

First, we show the score percentages of the participants in the two languages.

Table 1 shows the mean scores for both sexes are lower for Turkish, which reflects the fact that this test was more difficult (included more infrequent words). To test whether the data are normally distributed, we computed the Sapiro–Wilk test in SPSS. For both sexes, the scores are significantly different from a normal distribution at $p < 0.001$ for Turkish and for English at $p < 0.01$. For both languages, the distribution of the data is left-skewed and higher scores are more frequent than lower scores. We therefore use only non-parametric tests in the following computations.

One of the main questions of this present study is the relationship between the vocabulary sizes of male and female partners. Table 2 shows the correlations between the vocabulary sizes as measured with the Turkish and the English X-lex tests for the partners (between and within participants).

Table 1. Test scores (in percentages, n = 83).

	Min	Max	Mean	St.Dev
Male Eng	55	94	79.6	10.23
Male Turk	40.67	96.67	65.71	18.36
Female Eng	56	94	78.8	9.98
Female Turk	40	98	65.73	18.59

Table 2. Correlations (Spearman) between vocabulary sizes (n = 83).

	Male English	Male Turkish	Female English	Female Turkish
Male English	-	−0.722 **	0.956 **	−0.705 **
Male Turkish		-	−0.702 **	0.950 **
Female English			-	−0.672 **
Female Turkish				-

* = 0.05; ** = 0.01 (two tailed).

All correlations are significant with rho = 0.956 for the vocabulary sizes of the couples in English, and rho = 0.950 for the vocabulary sizes in Turkish. There is a reverse picture for the vocabulary sizes within participants in both languages. Participants with a higher vocabulary in English have lower vocabulary sizes in Turkish and vice versa. This is in line with the expectation that the languages of bilinguals follow a complementary principle (Grosjean [1982] 2001).

In the following, we computed partial correlations with controlling for the years of marriage to show whether accommodation during the length of marriage is the reason for the high correlations between the partners (see Table 3).

Table 3. Partial correlations (Spearman) between the scores of males and females (controlling for the years of marriage, n = 83).

	Male English	Male Turkish	Female English	Female Turkish
Male English	-	−0.822 **	0.959 **	−0.817 **
Male Turkish		-	−0.812 **	0.991 **
Female English			-	−0.798 **
Female Turkish				-

* = 0.05; ** = 0.01 (two tailed).

The correlations are almost similar to the correlations in Table 2. This means that it is not accommodation during the marriage that leads to similar vocabulary sizes, but that vocabulary played a crucial role at the time of the partner selection, probably in a very short period. This also indicates that humans can estimate the vocabulary size of an interlocutor unconsciously as the partners certainly did not administer a formal language test when they first met.

The findings are significant, which means that they are generalizable towards the population. However, as always in inferential statistics, you have to argue what the population might be but you cannot show what it is mathematically. To attain some further insight into a possible wider population, we rearranged the data by reallocating the partners of our sample on a random basis. If this leads to similar findings as the original data, we are more confident that our results are not just restricted to a narrow population.

As expected, the correlations between the vocabulary sizes of these randomly selected pairs are not significant (rho = 0.025 for English, and rho = 0.056 for Turkish), as shown in Table 4. Only the within participants correlations are strongly negative, as explained before.

The non-significant correlations between re-arranged partners are a clear indication that it is the vocabulary sizes of the original real partners that are relevant for partner selection.

Table 4. Spearman correlations between vocabulary sizes with re-arranged partners (within and between participants, n = 83).

	Male English	Male Turkish	Female English	Female Turkish
Male English	-	−0.722 **	−0.025 n/s	0.038 n/s
Male Turkish		-	−0.038 n/s	−0.056 n/s
Female English			-	−0.688 **
Female Turkish				-

* = 0.05; ** = 0.01 (two-tailed).

6. Discussion and Conclusions

From the literature review, it is evident that intelligence, education, and wealth are important assets for the value of a person on the biological market of partner selection. These psychological and economic traits cannot be judged directly but need to be inferred by a proxy, such as the vocabulary size. Vocabulary size cannot be manipulated at will and is therefore an honest signal for these traits. The format of our language tests provides valid information about participants' vocabulary sizes because of corrections for guessing. In other words, the test forces participants to use an honest signal only. Our findings show that there is a high correlation between the test scores of our participants and thus their vocabulary sizes. Therefore, (alternative) Hypothesis 1 is supported, which assumes high correlations in the test scores and that vocabulary is a proxy for the market value of potential partners. Of course, one has to bear in mind that correlations are not necessarily causations, and therefore our conclusions in general have to be taken cum grano salis.

In our sample, it is not the male or the female partner that has a larger vocabulary, but both partners have equal sizes with very high correlations between the partner that approach the maximum value for correlations. This supports (alternative) Hypothesis 2, that it is indeed a mutual decision made on a biological market where the best mate is chosen according to the "value" of a participant and that selections are made between likes, that is between people with a similar "value" on the market. The specific aspect of our sample is that the participants are sequential bilinguals with a complementary distribution of their vocabulary knowledge according to (Grosjean [1982] 2001). Indeed, there are high negative correlations between the two languages among participants (Table 2), which is in line with the complementary principle. However, there are high correlations between participants in both languages (Table 2), which indicates that both languages play a positive role in the display of the market value of our participants, which supports (alternative) Hypothesis 3. This is an indication that there are psychological or other traits underlying both languages (see Grosjean [1982] 2001; Daller and Ongun 2018) although this assumption cannot be proven directly. An important question in partner selection is how quickly decisions are made (see research question 3).

We partialled out the years of marriage in our sample to find out whether accommodation during marriage leads to similar vocabulary sizes in the partners or whether the vocabulary sizes already played a crucial role at the time of partner selection. Table 3 shows that years of marriage or time do not play a role in the similarities in vocabulary sizes between the partners. Independent of the years of marriage, the correlation between the vocabulary sizes in Table 3 is as high as in Table 2. Alternative Hypothesis 4 is therefore supported, and it is most likely that the judgement on the characteristics of a potential partner is made very quickly. How quickly this judgment is transferred into a real decision for partner selection will of course depend on various other variables that our out of the focus of this present study, e.g., potentially the influence of parents in different cultures. We are aware that these other factors, including belonging to similar social groups and many more, might be important for the decision-making process, but it is difficult to quantify

this without in-depth background information. This is certainly a limitation of this present study but does not contradict our conclusion that similarities in vocabulary sizes are a factor in partner selection.

In marriages that are arranged by the parents, this selection process might be slightly different, but still, the vocabulary size and the market value of a potential partner will be important. Finally, we rearranged our couples and allocated females to the male partners at random. The picture is clear from Table 4 that in this case, no significant correlations can be found and that the original not re-allocated partnerships are relevant. This gives us some indication that our findings are not only valid for our specific sample but to a potential wider population.

Overall, to our knowledge, we show for the first time how vocabulary size is important in marital partner selection. This is also in support of the assumption in evolutionary psychology (see Miller 2000) that the development of a large vocabulary is an advantage in mate selection. Miller argues that "verbal courtship is mutual" (Miller 2000, p. 383), and that is exactly what we find in our data—it is both males and females who display their market value through vocabulary on the biological market of partner choice. Although our findings are only based on middle-class participants, they support Miller's theoretical framework on human evolution in general. Our findings might therefore be generalizable in a wider context than originally thought.

A further limitation of our study is that we measured vocabulary size only with a receptive vocabulary test. However, there are strong correlations between productive and vocabulary knowledge (see Ha and Nguyen 2023; Milton and Fitzpatrick 2017; Roghani and Milton 2017; Webb 2008; Zhang and Zhang 2022) and we therefore accept that our test can be used as a proxy in the present research setting. A larger test battery would have been desirable and might be used in further studies.

We acknowledge that it is a massive leap to draw conclusions from the present correlations about a decision-making process years ago when the partners met. However, we still claim that knowledge of infrequent vocabulary, be it productive or receptive, is human capital that is acquired through a lengthy and costly process of education. It is of course only one asset amongst many others that are important in partner selection. A further limitation of our study is certainly the fact that we have only data from a certain socio-economic background, and it might be the case that other assets that are not so closely related to education as vocabulary are more important in other settings. We agree that our findings are just a first step for future research in partner selection that certainly will have to include many other variables, which would then allow researchers to sketch a more fine-grained picture of the processes that are involved in partner selection.

Author Contributions: Conceptualization, M.D. and Z.O.; formal analysis, M.D.; investigation, Z.O.; writing—original draft preparation, M.D.; writing—review and editing, M.D.; project administration, Z.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The project met all the requirements for ethical control in place in Swansea University. An approval number was not required at the time.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be requested from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 We used SPSS version 27 to carry out a partial correlation (command: correlate > partial . . . controlling for . . ."). This removes the influence of the variable that is not of interest ("years of marriage") in the model (see also Aloe and Thompson 2013).

2 In principle, it might be possible that the market value of the females is more important than the market value of the males, but there is no indication from the literature that this might be an option (see Miller 2000), and we did not consider this possibility so as to keep our design as simple as possible.

References

Ahern, Kenneth R. 2017. Information networks: Evidence from illegal insider trading tips. *Journal of Financial Economics* 125: 26–47. [CrossRef]

Aloe, Ariel M., and Cristopher G. Thompson. 2013. The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research* 4: 390–405. [CrossRef]

Barclay, Pat. 2013. Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior* 34: 164–75. [CrossRef]

Barclay, Pat. 2016. Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology* 7: 33–38. [CrossRef]

Becker, Gary S. 1976. *The Economic Approach to Human Behavior*. Chicago and London: University of Chicago Press.

Becker, Gary S. 2007. Health as human capital: Synthesis and extensions. *Oxford Economic Papers* 59: 379–410. [CrossRef]

Becker, Gary S. 2009. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press.

Becker, Gary S., Julio J. Elias, and Kevin M. Murphy, eds. 2023. *The Economic Approach: Unpublished Writings of Gary S. Becker*. Chicago: University of Chicago Press.

Bourdieu, Pierre. 2018. The forms of capital. In *The Sociology of Economic Life*. London: Routledge, pp. 78–92.

Brown, Dale. 2010. *Brian Richards, Michael H. Daller, David D. Malvern, Paul Meara, James Milton and Jeanine Treffers-Daller: Vocabulary Studies in First and Second Language Acquisition*. London: Palgrave Macmillan.

Buss, David M., M. Abbott, A. Angleitner, A. Asherian, A. Biaggio, A. Blanco-Villasenor, M. Bruchon-Schweitzer, H. Y. Ch'u, J. Czapinski, and B. Deraad. 1990. International preferences in selecting mates: A study of 37 cultures. *Journal of Cross-Cultural Psychology* 21: 5–47. [CrossRef]

Clark, Eve V., and Marisa Casillas. 2015. First language acquisition. In *The Routledge Handbook of Linguistics*. London: Routledge, pp. 311–28.

Coleman, James S. 1993. The impact of Gary Becker's work on sociology. *Acta Sociologica* 36: 169–78. [CrossRef]

Cummins, James. 1976. The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism* 9: 1–43.

Cummins, James. 1991. Interdependence of First- and Second-Language Proficiency in Bilingual Children. In *Language Processing in Bilingual Children*. Edited by Ellen Bialystok. Cambridge: Cambridge University Press, pp. 70–89.

Daller, Michael, and Zehra Ongun. 2018. The threshold hypothesis revisited: Bilingual lexical knowledge and non-verbal IQ development. *International Journal of Bilingualism* 22: 675–94. [CrossRef]

Deane, Kevin. 2020. Can economics explain everything and solve all our problems? In *Recharting the History of Economic Thought*. Edited by Kevin Deane and Elisa Van Waeyenberge. London: Red Globe Press, pp. 287–302.

Deary, Ian J., S. Strand, P. Smith, and C. Fernandes. 2007. Intelligence and educational achievement. *Intelligence* 35: 13–21. [CrossRef]

Delcey, Thomas. 2019. Samuelson vs Fama on the efficient market hypothesis: The point of view of expertise. *Œconomia. History, Methodology, Philosophy* 9-1: 37–58. [CrossRef]

Devenport, Scott, Catriona Davis-McCabe, and Sam Winter. 2023. A Critical Review of the Literature Regarding the Selection of Long-Term Romantic Partners. *Archives of Sexual Behavior* 52: 3025–42. [CrossRef] [PubMed]

Dunbar, R. I. M. 2001. The economics of male mating strategies. *Economics in Nature: Social Dilemmas, Mate Choice and Biological Markets* 2001: 245–69.

Dupont, Emelien, A. Van Pottelberge, B. Van de Putte, J. Lievens, and F. Caestecker. 2020. Divorce in Turkish and Moroccan communities in Belgium. *European Journal of Population* 36: 617–41. [CrossRef]

Eisenbruch, Adar B., and James R. Roney. 2017. The skillful and the stingy: Partner choice decisions and fairness intuitions suggest human adaptation for a biological market of cooperators. *Evolutionary Psychological Science* 3: 364–78. [CrossRef]

Fisman, Raymond, S. S. Iyengar, E. Kamenica, and I. Simonson. 2006. Gender differences in mate selection: Evidence from a speed dating experiment. *Quarterly Journal of Economics* 121: 673–79. [CrossRef]

Frieder, Rachel E., C. H. Van Iddekinge, and Patrick H. Raymark. 2016. How quickly do interviewers reach decisions? An examination of interviewers' decision-making time across applicants. *Journal of Occupational and Organizational Psychology* 89: 223–48. [CrossRef]

García-Carpintero, Manuel. 2023. Lying versus misleading, with language and pictures: The adverbial account. *Linguistics and Philosophy* 46: 509–32. [CrossRef]

Geher, Glenn, S. B. Kaufman, J. A. Planke, and J. M. Di Santo. 2020. Mating intelligence. In *The Cambridge Handbook of Intelligence*. Edited by R. J. Sternberg. Cambridge: Cambridge University Press, pp. 846–70.

Grosjean, François. 2001. *Life with Two Languages: An Introduction to Bilingualism*. Cambridge: Harvard University Press. First published 1982.

Grosjean, François. 2015. The Complementarity Principle and its impact on processing, acquisition, and dominance. In *Language Dominance in Bilinguals: Issues of Measurement and Operationalization*. Edited by Carmen Silva-Corvalan and Jeanine Treffers-Daller. Cambridge: Cambridge University Press, pp. 66–84.

Ha, Hung Tan, and Duyen Thi Bich Nguyen. 2023. Using structural equation modelling to examine the relationship between receptive vocabulary knowledge and receptive language proficiency: A new light on an old issue. *International Journal of Applied Linguistics* 33: 382–98. [CrossRef]

Haandrikman, Karen, and L. J. van Wissen. 2012. Explaining the flight of cupid's arrow: A spatial random utility model of partner choice. *European Journal of Population/Revue Européenne de Démographie* 28: 417–39. [CrossRef]

Heckman, James J., Edward P. Lazear, and Kevin M. Murphy. 2018. Gary Becker Remembered. *Journal of Political Economy* 126: 1–6. [CrossRef]

Hegelund, Emelie Rune, M. Grønkjær, M. Osler, J. Dammeyer, T. Flensborg-Madsen, and E. L. Mortensen. 2020. The influence of educational attainment on intelligence. *Intelligence* 78: 101419. [CrossRef]

Hindmarsh, Roland. 1980. *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

Hodgeson, Geoffrey M. 2003. The hidden persuaders: Institutions and individuals in economic theory. *Cambridge Journal of Economics* 27: 159–75. [CrossRef]

Hodgeson, Geoffrey M. 2012. On the limits of rational choice theory. *Economic Thought* 1: 2012.

Kalmijn, Matthijs. 1991. Shifting boundaries: Trends in religious and educational homogamy. *American Sociological Review* 56: 786–800. [CrossRef]

Lavoie, Marc. 1992. *Foundations of Post-Keynesian Economic Analysis*. Cheltenham and Camberly: Edward Elgar Publishing.

Luxen, Marc F., and Bram P. Buunk. 2006. Human intelligence, fluctuating asymmetry and the peacock's tail: General intelligence (g) as an honest signal of fitness. *Personality and Individual Differences* 41: 897–902. [CrossRef]

Meara, Paul, and Antònia Babi. 2001. Just a few words: How assessors evaluate minimal texts. *IRAL-International Review of Applied Linguistics in Language Teaching* 39: 75–83. [CrossRef]

Meara, Paul, and James Milton. 2003. *X_Lex, the Swansea Levels Test*. Wolverhampton: Newbury Express.

Miller, Geoffrey. 2000. *The Mating Mind*. London: Random House.

Milton, James, and Jeanine Treffers-Daller. 2013. Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review* 4: 151–72. [CrossRef]

Milton, James, and Tess Fitzpatrick. 2017. *Dimensions of Vocabulary Knowledge*. London: Bloomsbury Publishing.

Nascimento, Bruna, Lora Adair, and Katia Vione. 2024. Pathways to online infidelity: The roles of perceived online dating success, perceived availability of alternative partners, and mate value discrepancy. *Current Psychology* 43: 12782–93. [CrossRef]

Nation, Paul. 1984. *Vocabulary Lists: Words, Affixes, and Stems*. Wellington: English University of Wellington. San Francisco: English Language Institute.

Nebl, P. J., M. G. McCoy, G. C. Foster, and M. J. Zickar. 2021. Assessment of the Mate Retention Inventory-Short Form Using Item Response Theory. *Evolutionary Psychology* 19: 1. [CrossRef]

Noë, Ronald, and Peter Hammerstein. 1995. Biological markets. *Trends in Ecology & Evolution* 10: 336–39.

Noë, Ronals, Jan Aram van Hooff, and Peter Hammerstein. 2001. *Economics in Nature: Introduction*. Cambridge: University Press, p. ix.

Pawlowski, Boguslaw, and Slawomir Koziel. 2002. The impact of traits offered in personal advertisements on response rates. *Evolution and Human Behavior* 23: 139–49. [CrossRef]

Pentland, Alex. 2010. *Honest Signals: How They Shape Our World*. Cambridge: MIT Press.

Ranzine, Giulia, and Cristoph Lutz. 2017. Love at first swipe? Explaining Tinder self-presentation and motives. *Mobile Media & Communication* 5: 80–101.

Roghani, Shadan, and James Milton. 2017. Using Category Generation Tasks to Estimate Productive Vocabulary Size in a Foreign Language. *TESOL International Journal* 12: 143–59.

Rosenberg, Jeremy, and Richard J. Tunney. 2008. Human vocabulary use as display. *Evolutionary Psychology* 6: 538–49. [CrossRef]

Scott, Jarvis. 2000. Rational choice theory. In *Understanding Contemporary Society: Theories of the Present*. Thousand Oaks: Sage Publications, vol. 129, pp. 671–85.

Smith, Kristopher M., and Coren L. Apicella. 2020. Partner choice in human evolution: The role of cooperation, foraging ability, and culture in Hadza campmate preferences. *Evolution and Human Behavior* 41: 354–66. [CrossRef]

Stamenković, Dušan, Nicholas Ichien, and Keith J. Holyoak. 2019. Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language* 105: 108–18. [CrossRef]

Taylor, Charles. 1994. *Multiculturalism*. Princeton: Princeton University Press.

Thorsteinson, Todd J. 2018. A meta-analysis of interview length on reliability and validity. *Journal of Occupational and Organizational Psychology* 91: 1–32. [CrossRef]

Toma, Catalina L., and Jeffrey T. Hancock. 2010. Looks and lies: The role of physical attractiveness in online dating self-presentation and deception. *Communication Research* 37: 335–51. [CrossRef]

Treffers-Daller, Jeanine, and Carmen Silva-Corvalán, eds. 2016. *Language Dominance in Bilinguals: Issues of Measurement and Operationalization*. Cambridge: Cambridge University Press.

Van der Zanden, Tess, Alexander Peter Schouten, M. B. Mos, and E. J. Krahmer. 2020. Impression formation on online dating sites: Effects of language errors in profile texts on perceptions of profile owners' attractiveness. *Journal of Social and Personal Relationships* 37: 758–78. [CrossRef]

Webb, Stuart. 2008. Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition* 30: 79–95. [CrossRef]

Zahavi, Amotz, and Avishag Zahavi. 1999. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.
Zhang, Songshan, and Xian Zhang. 2022. The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research* 26: 696–725. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

TED Talks and the Textbook: An In-Depth Lexical Analysis

Naheen Madarbakus-Ring [1,*] and Stuart Benson [2]

[1] CEGLOC, Institute of Humanities and Social Sciences, University of Tsukuba, Tsukuba 305-8577, Japan
[2] Center for Language Research, University of Aizu, Aizuwakamatsu 965-0006, Japan; benson@u-aizu.ac.jp
* Correspondence: drnmring@gmail.com

Abstract: The development of TED Talks textbooks has been a welcoming addition to English for Academic Purposes (EAP) pedagogy. The textbooks offer educators and learners a suitable framework for practicing all four of the language skills (i.e., listening, reading, speaking, and writing). However, the use of TED Talk resources could create specific vocabulary challenges for learners as they progress through each unit in the textbook. Research suggests that although textbook frameworks encompassing listening resources benefit learners with a familiar lesson approach, the varying vocabulary load and the presence of academic vocabulary and multiword units (MWUs) presented between the chosen resources and the textbook itself could lead to comprehension difficulties for learners. This study investigates the vocabulary of 12 TED Talks included in the commercial textbook Keynote 2 to understand the lexical profile, vocabulary load, and the academic and multiword unit coverage for each of the chosen listening texts. The results showed that the TED Talks selections and the textbook provided inadequate vocabulary practice, limited academic vocabulary exposure, and a lack of item repetition for learners. The study suggests the inclusion of ideal supplementary materials and appropriate TED Talk selections to help provide educators with suitable guidance to support their learners' varying vocabulary knowledge.

Keywords: word lists; vocabulary knowledge; textbooks; TED Talks; listening

1. Introduction

In English Language Teaching (ELT), the textbook provides educators with suitable frameworks, guidance, and language when teaching English for Academic Purposes (EAP) courses. Specifically, the textbook provides learners with a primary source of vocabulary that is integrated into the textbook content to help them maximize their learning. Sun and Dang (2020) noted the potential of using textbook lessons to help learners notice and evaluate the most useful words needed for their language learning. Recently, the development of commercial textbooks using resources such as TED Talks (i.e., Keynote and World English) has emphasized the need to understand the lexical demands of these standalone resources. Given that textbooks provide well-developed and structured instruction for educators, it remains unknown whether these additional resources are lexically appropriate for learners when completing the units' listening activities.

L2 research has now turned to investigating the suitability of authentic resources in listening to understand the chosen input included in various textbooks. Morrow (1977) defined authentic resources as "a stretch of real language produced by a real speaker or writer for a real audience and designed to carry a real message of some sort" (p. 270). As a result, publishers have commissioned textbook writers to include authentic resources in their selections. For example, Cengage Learning has published five textbook titles, which include TED Talks as the language input (see http://ngl.cengage.com/ted, accessed on 10 September 2024). Elk (2014) advocated the suitability of TED Talks as a listening resource for learners, drawing on the similarities between the lecture-style components of lecturers and presenters and their use of PowerPoint visuals to explain concepts. Given the

popularity of these resources, it is surprising that few studies have examined the suitability of TED Talks for listening input.

Specifically, vocabulary in authentic resources could be problematic for learners. Schmitt and Schmitt (2014) highlighted the importance of measuring the lexical coverage of items by their frequency bands (i.e., high/mid/low frequency). In turn, recent studies (see Benson and Madarbakus-Ring 2021; Sun and Dang 2020; Yang and Coxhead 2020) have used Nation's (2012) BNC/COCA 25,000 lists to analyze vocabulary by its frequency level. Similarly, studies using the academic word list (AWL) (Coxhead 2000) have investigated academic vocabulary in TED Talks, with results indicating a suggested coverage of 4% (see Coxhead and Walls 2012; Nurmukhamedov 2017). However, few studies have considered the representation of multiword units (e.g., lexical chunks that have one or more words) in texts. Identifying the potential difficulty of such lexical items, categorized by their generality (i.e., high-/mid-/low-frequency items) or specialism (i.e., academic, technical, or multiword units) in texts, could help teachers provide learners with the tools they need to address their comprehension difficulties more successfully. Therefore, this study investigates the lexical profile, vocabulary load, academic, and multiword unit coverage of the 12 chosen TED Talks used in the Cengage Learning textbook, Keynote 2. The analysis conducted considers the lexical frequency, vocabulary difficulty, and the coverage of different word types (i.e., academic and multiword units), and examines the suitability of these resources for vocabulary learning when listening to the chosen TED Talks in textbooks.

2. Literature Review

Previous studies have investigated the vocabulary load (Sun and Dang 2020), word frequency (Matsuoka and Hirsh 2010), and vocabulary knowledge (Yang and Coxhead 2020) that learners need to understand vocabulary input in commercial textbooks. A common approach to investigating the vocabulary load in textbooks is using West's (1953) General Service List (GSL). The GSL word list consists of the first 2000 word families to help researchers identify high-frequency words in a text. Schmitt and Schmitt (2014) also recognized the need to identify mid-frequency and low-frequency words, suggesting broader investigations that examine the lexical coverage of texts as needed. As a result, studies have used Nation's (2012) British National Corpus/Corpus of Contemporary American English (BNC/COCA) word lists that contain the 25,000 most frequent words for more comprehensive investigations of high-/mid-/low-frequency words in textbooks. From these approaches, understanding the lexical coverage and the word frequency included in textbooks can help to support language learners.

2.1. Vocabulary Load in Textbooks

Nation (2006) describes the lexical coverage in texts as "the percentage of running words in the text known by learners" (p. 61). Numerous studies (see Benson and Madarbakus-Ring 2021; Laufer 1992; Van-Zeeland and Schmitt 2013) have investigated the vocabulary load of texts using two lexical coverage thresholds: 95% and 98% coverage. In listening activities, Van-Zeeland and Schmitt (2013) noted that a 95% threshold indicates "good but not necessarily complete" comprehension for learners, while a 98% threshold represents "very good comprehension" (p. 475). In textbooks specifically, five studies used Nation's (2012) BNC/COCA 25,000-word list to investigate the vocabulary load to understand their lexical coverage. The results are shown in Table 1.

Table 1. Previous studies examining the vocabulary load of commercial textbooks (adapted from Benson and Madarbakus-Ring 2021).

Study	Textbook	95% Coverage	98% Coverage
Matsuoka and Hirsh (2010)	New Headway (Upper Intermediate)	2000	-
Hajiyeva (2015)	University Textbook (11 titles)	3500	9000
Coxhead et al. (2017)	ESP Textbooks (15 titles)	3000	7000
Sun and Dang (2020)	Yilin (Four levels)	3000	9000
Yang and Coxhead (2020)	New Concept English (Two levels)	3000 4000	5000 6000

Table 1 shows the levels of comprehension learners need in their knowledge of word families (i.e., all words related to the root word are counted as one) to understand each of the commercial textbooks. Warnby (2023) acknowledged that although previous studies have fluctuated to show varying thresholds between 86.7% and 96.7%, an understanding of between 95% and 98% of words is frequently chosen as the essential threshold to measure learners' vocabulary knowledge. Specifically, all of these studies suggest that 95% and 98% coverage thresholds show that learners need knowledge of between 2000 and 4000 word families for 95% comprehension of the lexical coverage. Although Hajiyeva's (2015) analysis of university textbooks and Yang and Coxhead's (2020) New Concept English analysis showed that learners required higher lexical coverage thresholds of around 4000 word families, understanding the requirements between these lower threshold levels helps teachers to support learners that possess varying vocabulary knowledge. Considering the 98% lexical coverage, the same studies showed a broader variation in reaching this higher threshold. Yang and Coxhead (2020) found a lower threshold of 5000–6000 word families was needed for the New Concept English titles. Sun and Dang's (2020) study on the Yilin series and Hajiyeva's (2015) university textbook showed that knowledge of the first 9000 word families is needed to achieve 98% comprehension. More surprisingly, Matsuoka and Hirsh (2010) found that the New Headway (Upper Intermediate) textbook content never reached 98% lexical coverage. These results indicate that learners need a broad range of vocabulary knowledge when using commercial textbooks irrespective of the assigned proficiency level, suggesting that further research is needed to understand the different lexical demands of individual textbooks required by learners.

2.2. Vocabulary Frequency in Textbooks

Another useful approach is understanding the vocabulary frequency included in commercial textbooks. Schmitt and Schmitt (2014) divided vocabulary frequency into three categories, high frequency, mid-frequency, and low frequency, which can be defined as follows (Schmitt and Schmitt 2014):

High frequency: First, second, and third 1000 word families (e.g., *go*, *buy*, and *watch*).
Mid-frequency: Fourth to eighth 1000 word families (e.g., *academic*, *frequent*, and *octopus*).
Low frequency: Ninth group upwards (e.g., *outlandish*, *florescent*, and *azalea*).

As defined above, the vocabulary frequency bands divide Nation's (2012) BNC/COCA 25,000 into 3 categories that can help better understand learners' required vocabulary knowledge. From these 3 categories, Macalister and Nation (2020) commented on the importance of choosing texts that have high-frequency and wide-ranging input for learners. Specifically, Yang and Coxhead (2020) also described how these content and function words are usually the majority of high-frequency words included in English texts and, consequently, will be the words that learners identify the most. Although various studies have investigated high-frequency words in texts (see Table 1), Schmitt and Schmitt (2014) suggested that further research is needed to understand the mid-frequency and low-frequency word requirements (i.e., by creating word lists) to facilitate supporting learners' comprehension. The importance of providing opportunities for learners to acquire more difficult and complex mid-frequency and low-frequency words was also noted by Mo and

Bi (2024). For example, knowledge of mid-frequency words can often support learners in their reading fluency and, while knowledge of low-frequency words is rarely needed, as these items are infrequently used in the text, identifying such items can determine the potential vocabulary difficulties that may exist for learners (Yang and Coxhead 2020). Sun and Dang (2020) advocated how categorizing Nation's (2012) BNC/COCA lists into the 3 high-/mid-/low-frequency categories provides a more representative analysis of the lexical difficulties that learners may encounter. Learners need to be exposed to more repeated encounters with complex and difficult words that occur in the mid-frequency and low-frequency categories for learners to acquire knowledge of these words (Mo and Bi 2024). In turn, identifying these word frequency categories can help teachers to understand the most appropriate words to teach learners. However, categorizing words into high-/mid-/low-frequency lists can often be time-consuming for teachers, as the terms are rarely presented using these divisions in textbooks.

Further, studies investigating lexical profiles by word frequency have seldom investigated solely listening input in commercial textbooks. For example, O'Loughlin's (2012) study investigated the listening texts in the New English File textbook series. He analyzed 29,716 running words from the series' Elementary, Pre-intermediate, and Intermediate textbooks. Although he found that the vocabulary levels increased in difficulty for each level, he also found a variance in the lexical demands of the listening resources included in each textbook. Specifically, O'Loughlin (2012) found that out of a total of 1045 word families used across all 3 levels of the textbook, only 860 of these word families were used in the Intermediate textbook. These results indicate that as the 185 word families included in the first 2 textbook levels (i.e., Elementary and Pre-intermediate) were not included in the Intermediate textbook, learners may find these items difficult to master. This finding indicates the importance of repeating lexical terms throughout different levels of a textbook series. Specifically, learners need to master appropriate levels of vocabulary (i.e., low/mid/high) before studying the next level of the textbook. However, as this is not always possible, repetition of previous lexical items that feature in a textbook series is needed to enable learners to master words from previous textbooks when studying with higher levels of the series. In this example, O'Loughlin (2012) explained how learners may have lexical difficulties when using higher levels of the English File textbook if they do not have the opportunity to master these word families in the first two levels. As Mo and Bi (2024) further explained, more sophisticated words were not included beyond the first 1500 most frequent words, suggesting that words from the higher second and third lists were underrepresented or lacked repeated exposure. Therefore, these findings indicate the importance of repeating lexical items in textbooks and using relevant resources to support learners' vocabulary knowledge, in line with their textbook and learning progression. If words from previous textbooks are not included in later titles, teachers may have to rely on the assumption that learners have achieved a certain vocabulary threshold, resulting in problematic vocabulary for them if they meet unfamiliar lexis in higher textbook levels. Thus, in this study, examining the repetition of the lexical demands for each TED Talk, as used in a textbook series, can help educators to identify the lexical knowledge needed by learners for each sequential unit.

2.3. TED Talks and Vocabulary Learning

Vocabulary learning can be categorized into an academic word list (AWL), as devised by Coxhead (2000), to understand the occurrence of certain words in academic texts. Specifically, Coxhead's (2000) word list consists of 25 lists, each comprised of 1000 words (or a total of 25,000 words), drawing from 570 word families to present a comprehensive list, totaling 10% of the academic words used in academic texts. As learners progress through each subsequent list, the words become less frequently used in academic texts. Compared to Nation's (2012) BNC/COCA 25,000 list, AWL items tend to appear more often in academic rather than non-academic texts, providing a useful indicator of possible vocabulary difficulties for learners in academic contexts.

In academic learning, TED Talks are used as a listening resource in many classrooms (see Liu and Chen 2019; Mojgan and Tollabi 2019). TED Talks emerged as a global non-profit platform to showcase talks on more than 300 research topics listed under 6 categories: Business, Design, Education, Global Issues, Science, and Technology. Since 2007, TED Talks can be easily accessed for free from their website (www.ted.com accessed on 10 September 2024) for viewers to choose from both native and non-native expert presenters speaking about topics in more than 40 different languages. In English, both first-language and second-language experts speak in a range of Global Englishes, with talks accompanied by resources, such as transcripts and subtitles that can support L2 instruction. Romanelli et al. (2014) described the suitability of TED Talks in learning as shorter, unrestricted video accounts that help deliver lecture-type topics to learners. Further, TED Talks are often described as high-quality, professionally presented, and culturally relevant presentations that help learners digest a variety of topics. However, Field (2008) noted the generality of authentic resources that are "designed without language learning in mind" (p. 274). Therefore, understanding the vocabulary demands of authentic resources, such as TED Talks, could help determine their suitability for learners. To help ascertain the optimum lexical demands of TED Talks for language learning, previous studies have examined the vocabulary load of selected talks, as shown in Table 2.

Table 2. Vocabulary load findings from selected TED Talks.

Study	**Coxhead and Walls (2012) 60 Talks**		**Liu and Chen (2019) 2089 Talks**		**Nurmukhamedov (2017) 400 Talks**	
Threshold	95%	98%	95%	98%	95%	98%
All Talks	4000	9000	-	-	4000	8000
Business	5000	9000	-	-	4000	8000
Culture	-	-	3000	6000	-	-
Design	5000	8000	3000	6000	-	-
Entertainment	5000	8000	3000	6000	-	-
Global Issues	5000	9000	3000	5000	4000	8000
Science	5000	8000	3000	7000	4000	10,000
Technology	5000	9000	3000	6000	4000	8000

Note: the 95%/98% threshold was calculated according to Nation (2012).

As Table 2 shows, a variance in vocabulary knowledge is required by learners to understand these TED Talks. For example, all three studies showed that learners need vocabulary knowledge of between the first 3000 and 5000 word families to achieve 95% understanding. The higher threshold of 98% understanding requires learners to have vocabulary knowledge of between 5000 and 10,000 word families. Coxhead and Walls' (2012) smaller corpus of 60 TED Talks and Nurmukhamedov's (2017) larger corpus of 400 TED Talks both found that learners needed 4000–5000 word families for 95% coverage and 9000 word families for 98% coverage. Although Liu and Chen's (2019) larger corpus of 2089 TED Talks resulted in only 3000 word families for 95% coverage and 5000–7000 word families needed for 98% coverage, the results from all the studies concurred that learners need a large variance of between 5000 and 10,000 words to understand a range of TED Talks consisting of different topics and themes. These thresholds and potential lexical demands highlight the need to understand the individual vocabulary knowledge of the TED Talks used in commercial textbooks to prepare learners for their listening input. Although these previous studies indicate that vocabulary knowledge of between 3000 and 5000 word families may be optimum for Intermediate learners, further research is needed to understand the vocabulary demands, potential language difficulties, and necessary scaffolding for learners when listening to TED Talks included in Pre-intermediate commercial textbooks.

2.4. Academic Vocabulary in TED Talks

Understanding the academic vocabulary content in listening texts is another approach to measuring potential lexical difficulty for learners. For learners intending to study in academia, general academic vocabulary is considered important to learn, as it makes up approximately 9% of academic written texts (Nation 2022). Similarly, Warnby (2023) noted how academic vocabulary consists of about 10–15% of most written academic texts used for reading, suggesting that academic vocabulary knowledge is important for learners in university studies. As Warnby (2023) explained, dependent on the learners' vocabulary level, learners with lower vocabulary knowledge may find these reading texts, and their consequent lexical burden, more problematic. However, understanding academic vocabulary in listening texts is another approach to measure potential lexical difficulty for learners. Researchers (Coxhead and Walls 2012; Nurmukhamedov 2017; Wingrove 2022) have turned to investigating the academic vocabulary in TED Talks to help support learners in understanding the resource better in their English for Academic Purposes (EAP) courses. To date, four studies have examined academic vocabulary coverage using the AWL to analyze the vocabulary in TED Talks, which are shown in Table 3.

Table 3. Previous studies examining the academic vocabulary coverage of TED Talks.

Study	TED Talks Corpus	Word List Analysis	Academic Vocabulary Coverage
Nurmukhamedov and Sadler (2011)	Ken Robinson "Schools Kill Creativity" (1 TED Talk)	AWL	5%
Coxhead and Walls (2012)	Six by Six corpus (60 TED Talks)	AWL	3.90%
Nurmukhamedov (2017)	The TED Talks Corpus (400 TED Talks)	AWL	3.79%
Wingrove (2022)	TED Talks Corpus (2483 TED Talks)	AWL/AVL	4.09%

Note: all studies used the academic word list (AWL) (Coxhead 2000) for the vocabulary analysis.

As Table 3 shows, the four studies showed some agreement between academic vocabulary coverage found in TED Talks. In Nurmukhamedov and Sadler's (2011) analysis of one TED Talk, a higher academic coverage of 5% was found in Ken Robinson's "Schools Kill Creativity" talk. However, a more thorough analysis of larger TED Talks corpuses conducted in the three other studies (Coxhead and Walls 2012; Nurmukhamedov 2017; Wingrove 2022) showed similar findings of around 4% academic coverage. In line with Coxhead and Walls' (2012) initial findings of 3.90% in their analysis of 60 TED Talks, the results from these studies suggest that around 4% of the vocabulary in the TED Talk is academic. It is worth noting that different academic word lists were used in the four studies, three used the AWL (Coxhead 2000), while one study (Wingrove 2022) also used the academic vocabulary list (AVL) (Gardner and Davies 2014). While two lists were used, the coverage of academic vocabulary, at 4%, was similar. Wingrove's (2022) study highlighted discrepancies in coverage between academic word lists. When comparing the results of lexical analyses, the medium coverage varied from 3% to 4% (when using the AWL) to 17–18% (when using the AVL total list). To date, no study has used the updated NAWL in a lexical analysis of TED Talks. In listening, this observation highlights the importance of identifying the academic items within a TED Talk included in textbooks that could help support learners in their listening comprehension when completing the unit. This study aims to identify the academic coverage found in the specific Pre-intermediate TED Talks textbook using the NAWL, to provide learners with the appropriate lexical support for each unit.

2.5. Multiword Units in Textbooks and TED Talks

Multiword units (MWUs) is another crucial approach to understanding potential lexical demands in language learning (Ackermann and Chen 2013; Biber 2009). Defined as "phrases that are made up of words that frequently occur together" (Nation et al. 2016, p. 71), MWU is an umbrella term that incorporates various phrase types, including collocations (Siyanova and Schmitt 2008), idioms (Siyanova-Chanturia and Martinez 2015), lexical bundles (Byrd and Coxhead 2010), formulas (Simpson-Vlach and Ellis 2010), word clusters (McCarthy and Carter 2006), and formulaic sequences (Wood 2002). MWU sequences can be problematic for learners who may interpret each word individually rather than as a chunk of language. Wang and Liu (2023) suggested that increased encounters with both single words and collocations can result in higher gains of vocabulary knowledge for learners. However, in EAP written discourse, several studies have examined the coverage of MWUs (i.e., Ackermann and Chen 2013; Byrd and Coxhead 2010; Wood and Appel 2014), concluding that MWUs rarely occur in written texts and are not explicitly taught. In spoken discourse, Foster (2001) identified that speech is highly formulaic in general spoken discourse, while Erman and Warren (2000) noted that approximately 58.6% of lexicon consists of formulaic sequences. Few studies have investigated spoken MWUs (i.e., Coxhead et al. 2017), although Simpson-Vlach and Ellis' (2010) earlier study developed a spoken academic formulas list (AFL) containing 607 items based on a 2.1 million word spoken corpus of academic speech. However, to our knowledge, no study has investigated the coverage of MWUs in specific TED Talks. Therefore, this study aims to understand the inclusion of both general and academic MWUs of the selected TED Talks and their coverage in the textbook.

3. Materials and Methods

3.1. Research Questions

This study investigated the lexicon of TED Talks utilized in the commercial textbook, Keynote 2. The results will indicate whether the TED Talks included within a given textbook form a suitable resource for learners. The following research questions guided the study:

1. What are the vocabulary loads of the TED Talks used in Keynote 2?
   1a. How do the vocabulary loads of the TED Talks and Keynote 2 compare?
2. What is the lexical coverage of the academic words in the TED Talks used in Keynote 2?
   2a. How does the lexical coverage of the academic words in the TED Talks and Keynote 2 compare?
3. What are the multiword units (MWUs) that occur in the TED Talks used in Keynote 2?
   3a. How do the multiword units (MWUs) that coverage in the TED Talks and Keynote 2 compare?

3.2. TED Talks in Keynote 2

Keynote 2 (Bohlke et al. 2016) was first published by Cengage in 2015. The commercial textbook was chosen by the Introductory English course instructors at a public university in Japan. The textbook consists of 12 units that have a chosen TED Talk for each listening practice, which are described as suitable for Pre-intermediate learners (CEFR A2/B2). The 12 units were used to teach first-year university students over a 14-week semester. Each unit's TED Talk is supported by relevant content-based activities, which focus on each of the four skills (i.e., listening, reading, speaking, and writing). Each unit is divided into five lessons (i.e., Lessons A–E): Lessons A–C prepare learners about the TED Talk topic and content, Lesson D helps learners as they listen to the TED Talk, and Lesson E helps learners apply these ideas to real-world contexts. Additionally, each unit provides learners with topic-related vocabulary, visual aids, and content-based materials to help them to complete the textbook activities. Although Keynote 2 is only one textbook from a series of six, the aim of this study is to understand if the vocabulary in both the TED Talks and textbook correlate suitably for Pre-intermediate learners. As textbooks tend to be an integral tool for the teacher, the vocabulary content of the textbook is important to provide adequate

guidance for learners of a particular level in the classroom. Wang and Liu (2023) stated how materials are integral in prescribing the vocabulary knowledge to be taught for the teacher and the sequence and manner of the words to be presented to learners. Extensively analyzing one textbook to understand the vocabulary content of both the listening resources and the textbook content itself can help identify if authentic resources can be integrated into textbook materials for second language learners. Completing a thorough analysis of one textbook in the series can also help to indicate if the authentic resources and language content are level appropriate or if further analysis of the complete series is needed.

3.3. Data Collection

Keynote 2 was purchased and prepared for the data analysis. First, the TED Talk from each unit was tabulated and recorded by speaker, title, running time, and number of tokens (i.e., words). Each TED Talk is shown in Table 4.

Table 4. Keynote 2 listening texts by unit.

Unit	TED Talk	Listening Time	Number of Tokens
1	Munir Virani—"Why I Love Vultures"	6.03	954
2	A.J. Jacobs—"The World's Largest Family Reunion"	9.44	1466
3	Ann Morgan—"My Year Reading a Book from Every Country in the World"	12.04	1922
4	Daria van den Bercken—"Why I Take the Piano on the Road and in the Air"	9.31	582
5	Roman Mars—"Why City Flags May be the Worst Designed Thing You Have Ever Noticed"	18.09	3017
6	Jarrett Krosoczka—"How a Boy Became an Artist"	18.40	3283
7	Andras Forgacs—"Leather and Meat Without Killing Animals"	8.59	1149
8	Alessandra Orofino—"It Is Our City"	15.16	1902
9	Joy Sun—"Should You Donate Differently?"	7.35	1050
10	Tan Le—"A Headset that Reads Your Brainwaves"	10.31	1440
11	Louie Schwartzberg—"The Hidden Beauty of Pollination"	7.33	415
12	Nizar Ibrahim—"How We Unearthed Spinosaurus"	6.03	940
	Total	128.08 min/s	18,120 tokens

As the transcripts were not available in the students' textbook, the TED Talk transcripts were obtained from the official TED Talk website (www.ted.com accessed 10 September 2024) (TED Talks 2024). Each transcript was then cleaned using the following guidelines from Nation et al. (2016):

- Correct/check language to American English (memorise = memorize)
- Check spellings of unrecognized words
- Remove hyphens
- Remove apostrophes
- Delete spaces
- Rewrite contractions (haven't = have not)

The transcripts were first saved and checked as a Microsoft Word document (.docx file) and then saved and checked as a Notepad document (.txt file). As Nation et al. (2016) suggested, checking the data at least twice helps researchers to notice any missed spaces, apostrophes, or hyphens. The additional check also provides a further opportunity to add words that were not found in the BNC/COCA lists (Nation et al. 2016), such as proper nouns (e.g., Sakura and Woodford), compound words (e.g., basecamp and firewood), or newly coined terms that belong to a word family. If the items satisfy the criterion of levels three to six of Bauer and Nation's (1993) word family scale (i.e., most frequent regular derivational affixes, frequent and orthographically regular affixes, regular but infrequent affixes, and frequent but irregular affixes), the item can be added to the list. For example, the word *selfie* was added to BASEWRD1, as the meaning is related to the word family *self*. Nation et al. (2016) commented on the importance of these checks to continuously

update the BNC/COCA word list to ensure recency, validity, and reliability of lexical terms that are used in texts. To compare the vocabulary load in the TED Talks to each unit, the textbook was first scanned using Optical Character Recognition (OCR) software and cleaned following the guidelines above. Finally, 12 data files were created for each unit in the textbook.

3.4. Data Analysis

Each TED Talk transcript and unit was analyzed using Heatley et al.'s (2002) Range program and Nation's (2012) BNC/COCA 25,000 lists. The BNC/COCA word list is compiled of 25, 1000-word families from the British National Corpus (Ashton and Burnard 1998) and the Corpus of Contemporary American English (Davies 2008) and is ranked according to frequency and range. Additionally, Nation's (2012) four supplementary lists (i.e., proper nouns, marginal words, transparent compounds, and abbreviations) were also used in the analysis.

3.5. Corpus Analysis

AntWordProfiler (Version 2.2.1) (Anthony 2023), along with the new academic word list (NAWL) (Browne et al. 2013), were used to identify academic items in the TED Talks and textbook. The NAWL is a 957-word list of general academic items and was designed as an updated version to the original academic word list (AWL) by Coxhead (2000). While a spoken academic word list would be more applicable to identify items, currently, there is no lexical profiler that uses the academic spoken word list (Dang et al. 2017). Antwordprofiler (Anthony 2023) allows the user to compare texts and highlight items (in this case, the NAWL) that occur in each.

The multiword unit profiler (MWUP) (Version 2.0.0) (Eguchi 2021) was used to identify general and academic MWUs in the TED Talks. The phrasal expressions list (Martinez and Schmitt 2012) contains 505 of the most frequent non-transparent expressions in English. The academic formulas list (AFL) (Simpson-Vlach and Ellis 2010) contains 607 written and spoken MWUs from general academic contexts. When run through MWUP, it displayed frequency data for the identified items in each word list. Once analyzed, the relative frequency of each MWU was calculated using its frequency, divided by the number of running words, and multiplied by one million (i.e., $8 \div 18{,}120 \times 1{,}000{,}000$). For example, *part of the* occurred eight times and was divided by the TED Talks corpus (18,120 running words), then multiplied by one million to equal its relative frequency (=442). In the textbook, the relative frequency of *part of the* was 286 ($9 \div 31{,}462 \times 1{,}000{,}000$). As presented in the following Results Section, the vocabulary load analysis and the corpus analysis on both the TED Talks and the textbook units will indicate the lexical demands required by learners when using Keynote 2.

4. Results

This study examined the lexicon of the 12 chosen TED Talks and the textbook content in Keynote 2 using four single and MWU lists. Nation's (2012) BNC/COCA 25,000-word list was used to understand the vocabulary load and cumulative coverage of each TED Talk and textbook unit. Browne et al.'s (2013) NAWL identified 150 academic words in the TED Talks. Martinez and Schmitt's (2012) phrasal expressions list identified 141 general phrasal expressions, and Simpson-Vlach and Ellis' (2010) academic formulas list identified 97 academic MWUs. The following section shows the results for each of the three research questions.

4.1. Vocabulary Load of TED Talks in Keynote 2

Table 5 presents the cumulative coverage of Nation's (2012) BNC/COCA word lists for each TED Talk used in Keynote 2. As the supplementary lists (SL = proper nouns, marginal words, and compound words) are less of a learning burden compared to other words (Nation and Webb 2011), their coverage is shown at the top of the table for reference and then added accordingly. As illustrated in blue, 10 of the 12 units (Units 2, 3, 4, 5, 6,

7, 8, 9, 10, and 11) achieved 95% coverage between the 2000 and 3000 word families. As illustrated in red, the same units (except Units 4 and 11) also reached 98% between the 3000 and 5000 word families, suggesting that learners would need to know the first 5000 word families to understand a majority of the textbook's TED Talks. However, Units 1, 11, and 12 showed varying difficulty. Unit 12 results indicated that 95% coverage is reached at 4000 words and 98% coverage is achieved at 6000 words. Two units also indicated that a higher lexical range is needed, with 95% coverage achieved at 6000 (Unit 1) and 3000 (Unit 11) words and 98% reached at 8000 words for both units' talks, respectively. These results suggested vocabulary difficulty is likely, especially with the Unit 1 and Unit 11 TED Talks, as a broader vocabulary knowledge at the mid-level and low-level frequency is needed to understand the content.

Table 5. Vocabulary load of TED Talks across Nation's (2012) 25,000 BNC/COCA word lists.

	Unit												
Word List	**U1**	**U2**	**U3**	**U4**	**U5**	**U6**	**U7**	**U8**	**U9**	**U10**	**U11**	**U12**	**TOTAL**
SL 31–34	5.02	3.96	4.68	6.70	6.40	2.98	1.83	4.27	2.00	2.64	2.63	4.80	3.99
1	83.11	88.48	89.28	91.41	83.86	90.95	78.24	85.24	84.86	81.39	87.28	86.71	85.90
2	90.24	94.28	**94.69**	94.85	93.64	**95.64**	86.77	92.81	**94.86**	91.46	91.72	90.43	92.61
3	93.07	**97.76**	**98.18**	**96.57**	**96.89**	97.56	**95.91**	**97.96**	**98.19**	**95.49**	**95.56**	92.66	96.31
4	93.70	**98.37**	98.70	97.60	97.82	**98.47**	97.04	**98.59**	99.24	97.02	96.37	**95.21**	97.34
5	94.64	98.98	98.96	97.94	**98.42**	98.71	**98.43**	99.12	99.34	**98.48**	97.38	97.87	98.18
6	**95.37**	99.46	99.01	**99.49**	98.75	99.01	98.69	99.70	99.53	99.10	97.58	**98.19**	98.65
7	95.79	99.53	99.06	99.49	99.11	99.10	98.95	99.81	99.63	99.24	97.58	98.72	98.83
8	**98.73**	99.87	99.22	99.66	99.54	99.39	98.68	99.86	99.92	99.45	**99.20**	98.56	99.34

Note: Blue indicates 95% coverage and red indicates 98% coverage/SL = supplementary lists (31–34).

Table 6 compares the cumulative coverage of the combined 12 TED Talks scripts with the coverage of the other unit content in Keynote 2. The results showed that the TED Talks were slightly more lexically demanding than the textbook. A vocabulary size of 5000 word families is needed for 98% coverage of the TED Talks, compared to 4000 word families for the textbook. Although further analysis is needed to identify the specific vocabulary types to compare the lexical content in both the Keynote 2 unit and the TED Talks, the results showed that the textbook did not correlate the vocabulary knowledge with the lexical demands indicated in the TED Talks.

Table 6. Cumulative coverage of Nation's (2012) BNC/COCA lists in the TED Talks and Keynote 2.

Word List (By 1000 Words per List)	**Cumulative Coverage of TED Talks**	**Cumulative Coverage of Keynote 2**
Supp. lists	3.99	4.62
1st 1000	85.90	86.00
2nd 1000	92.61	94.47
3rd 1000	96.31	97.49
4th 1000	97.34	98.23
5th 1000	98.18	98.91
6th 1000	98.65	99.29
7th 1000	98.83	99.45
8th 1000	99.34	99.57
9th–25th 1000	100	100

Note: Blue indicates 95% coverage and red indicates 98% coverage; 9th–25th 1000 = word lists from the first 9000 to the first 25,000 lists.

4.2. Lexical Coverage of Academic Words in TED Talks Used in Keynote 2

Table 7 shows the frequency and coverage of the 150 academic words identified in the TED Talks from the NAWL (Browne et al. 2013). Additionally, Table 7 shows the items' frequency and their coverage in each unit of the textbook. The results highlighted three key points. First,

coverage of academic words in the TED Talks was low, with an average of only 1.24% from the 12 talks. Second, only 9 of the 150 items occurred in different TED Talks (i.e., *amongst*, *prey*, *sub*, *publish*, *essentially*, *reconstruct*, *afterwards*, *goods*, and *incredible*). This suggests that there was very little repetition of these items in the chosen audio forms. Finally, the academic words did not occur in the corresponding unit in the textbook. Only *ancestor* (7 times), *lab* (7 times), and *fossil* (10 times) occurred more than three times in the corresponding unit. The other items (see Table 7) seldom occurred in the unit. Therefore, this indicates that the textbook does not provide frequent opportunities for learners to encounter the academic words that they will hear in the TED Talks. Thus, the results indicate a lack of consistency in the academic vocabulary coverage found between the TED Talks and the textbook.

Table 7. Coverage of NAWL in each TED Talk and Keynote 2.

Unit	Lemma	Freq.	%	In TED Talks	Textbook %	Also Occurs in Unit (Frequency in Unit)
1	12	12	1.27	Amongst, flesh, ecological, bacteria, absorb, tremendous, unity, predator, critically, prey, ecology, locally	1.45	Predator (1) Flesh (2) Critically (2) Prey (2) Ecological (2)
2	11	13	0.89	Ancestor (s), offspring, continent, minus, migrate, accumulate, sub, sperm, donor, spectrum, tribe	0.62	Ancestor (7)
3	10	17	0.91	Publish (ed, ing), translation, facilitate, clarify, intensive, entities, manuscripts, rituals	0.64	Publish (2)
4	3	3	0.54	Composer, media, prejudice	0.71	Composer (1) Prejudice (1)
5	18	20	0.67	Meaningful, feedback, simplify, indicator, stripes, pole, essentially, horizontal, namely, loop, crude, painful, wheat, depict, reconstruct, founding, underneath, ridiculous,	0.97	Simplify (1) Meaningful (3)
6	15	25	0.76	Publish (ed, ing), monkey, classroom (s), maternal, elementary, afterwards, myth, cheated, blank, tempt, halfway, beaming, drained, consciousness, scholarship	0.64	Monkey (2) Publish (3) Classroom (2)
7	13	17	1.5	Lab (s), organs, essentially, goods, sophisticated, sustainable, technically, dimensional, horizon, multiply, matrix, insects, elasticity	1.83	Lab (7) Sustainable (2)
8	17	22	1.46	Incredible, incredibly, impact, consumption, emissions, aspects, sub, importantly, complement, collective, inequality, widespread, tech, authorities, allocate, amongst, afterwards	0.54	Consumption (1)
9	7	7	0.67	Economists, allocating, empirical, click, coordinates, corruption, goods	0.60	
10	23	35	2.47	Detection (s), algorithm (s), neurons, impulse (s), cortex, cognitive, neutral, interface, conscious, explicitly, realm, interact, emits, outer, identical, array, sensitivity, scroll, duration, analogies, goodness, differentiate, robots	1.68	Interface (1) Conscious (1)
11	5	5	1.05	Bats, colony, reproduce, reproduction, naked	0.71	Colony (1) Reproduction (2)
12	16	25	2.73	Incredible, fossil (s), bizarre, partial, dense, ultimate, prey, snakes, specimens, reconstruct, sediment, compact, limbs, fleshed, ruler, lifetime	1.49	Bizarre (2) Incredible (2) Fossil (10)
TOTAL	150	201	1.24		0.99	

4.3. Multiword Units Coverage of the TED Talks in Keynote 2

Table 8 presents the relative frequency per million of the 141 phrasal expressions and 97 academic formulas identified in the TED Talks and their relative frequency in the textbook. Upon closer analysis, eight phrasal expressions (e.g., *have to*, *a lot*, and *a few*) identified in the TED Talks had a relative frequency of over 501. When compared to the textbook, only 1 expression (*a lot*) had a relative frequency of over 501. This analysis indicates that coverage of the same MWUs in both the textbook and TED Talks could correlate more closely when designing activities in the curriculum design process. Similar to academic single words (see Table 7), Table 8 shows that there were fewer academic items (97) than general MWUs (141) included in the textbook. Therefore, the relative frequency of general and academic MWUs found in TED Talks was larger than those found in the textbook.

Table 8. Frequency per million of the phrasal expressions and academic formulas in each TED Talk and Keynote 2.

Relative Frequency	TED Talks	Textbook	Examples
0	0	76	I mean, those who, to do with
	0	60	you have a, you look at, we can see
1–50	0	28	take up, the case, as a result
	0	18	you need to, we look at, to look at
51–100	74	23	other than, a range of, as well
	57	13	they do not, to each other, there may be
101–150	26	5	in front of, choose to, so that
	20	2	we need to, the number of, at the same time
151–200	15	4	a couple of, find out, you see
	10	3	the end of, of the same, if you want to
201–250	9	1	of course, think that, no idea
	5	0	you have to, we need to, if you have
251–300	6	1	at least, find out, sort of
	1	1	you can see
301–350	3	1	look like, work on, right now
351–400	0	0	
	3	0	be able to, look at the, you want to
401–450	0	1	such a
	1	0	part of the
451–500	0	0	
501 and above	8	1	have to, a lot, a few

White—phrasal expressions (total MWUs—141); grey—academic formulas (total MWUs—97).

5. Discussion

The analysis showed that the TED Talks and the textbook provided inadequate practice for the other when developing vocabulary knowledge: the TED Talks are lexically more demanding than the textbook. In line with previous research, the TED Talks selected for this textbook had cumulative thresholds of 3000 word families for 95% and 5000 word families for 98% coverage (see Elk 2014; Liu and Chen 2019). However, there were several inconsistencies, with Units 1, 11, and 12 reaching the 95% threshold between 3000 and 6000 word families, and reaching the higher 98% threshold between 5000 and 8000 word families, which are similar to other studies (Coxhead and Walls 2012; Nurmukhamedov 2017). In contrast to Field's (2008) suggestions of language difficulty progressing systematically, these findings suggest that the TED Talks chosen for this textbook may not grow in difficulty sequentially. As the results illustrated, earlier units (e.g., Unit 1) are more difficult to understand than later units (i.e., Unit 4). The vocabulary level of TED Talks in textbooks should be monitored more closely to ensure that the level of difficulty develops in line with

the learners' expected language progress. As the topic of each unit is based on the TED Talks, it would seem that the talks are not a standalone resource and, therefore, the textbook should be preparing the learner to understand the talk by introducing critical vocabulary. However, in the current state, teachers would have to identify the vocabulary load of both the textbook and TED Talk, and prime learning of specific vocabulary items to bridge the gap between learners' vocabulary knowledge and the 8000 word family threshold needed to achieve 98% comprehension.

While studies have noted TED Talks can be appropriate learning resources (Liu and Chen 2019; Mojgan and Tollabi 2019; Romanelli et al. 2014), it is important to note that the resource was not designed for academia. As Field (2008) describes, authentic resources, such as TED Talks, were not designed primarily for language learning activities. These observations point to the limitation of using TED Talks in learning, as the lack of control over the vocabulary may not be suitable for second language learners. Thus, the results of this study indicate that more attention should be paid to analyzing the specific language needed by learners in TED Talk resources to use in correlation with textbook activities. TED Talks used in textbooks, such as Keynote 2, could better prepare learners by understanding the language used by the speaker, categorizing the items by frequency or word type (i.e., academic and multiword units), and focusing on repetition of these terms throughout the textbook activities. For example, introducing and repeating vocabulary that is key to the video could expose learners to the items they need to comprehend the TED Talk. Therefore, repeating this study's analysis of Keynote 2 with the other textbooks in the series could further ascertain the lexical needs that learners may need to pay more attention to in lessons when using authentic resources.

The corpus analysis also revealed two results. First, the TED Talks analysis contained few instances of academic vocabulary, with only 1.24% coverage from the NAWL (Browne et al. 2013). Further, in the textbook content itself, only 0.99% academic coverage of the items from the TED Talks was found. In both instances, the coverage was below previous findings of 4%, which could be a result of the analysis using the NAWL instead of the AWL (Coxhead 2000) and AVL (Gardner and Davies 2014) (see Table 3). In this study, the researchers deemed the NAWL as the most appropriate analytical tool because it is a larger and updated corpus of 288 million tokens (Browne et al. 2013), compared to the AWL corpus of 3.5 million tokens (Coxhead 2000). However, further research could consider whether 4% coverage (as suggested by studies analyzing academic vocabulary using the AWL) is appropriate or if an alternative NAWL threshold is needed. As the findings in this study indicated a lower NAWL coverage of 1.24%, the AWL coverage threshold of 4% may not be applicable in this instance. Further analyses using the NAWL (Browne et al. 2013) could identify the most appropriate coverage threshold for academic vocabulary in TED Talks.

Warnby (2023) also noted how the inclusion of academic vocabulary may correlate positively with general vocabulary to provide more exposure for learners to encounter these words. However, learners listening to the TED Talks in this textbook may find the listening selections are not providing adequate coverage of academic vocabulary in the corresponding unit content. These findings indicate that the selected TED Talks may not provide learners with the suitable learning opportunities that they need in the textbook to acquire the academic vocabulary included as their language learning progresses. Thus, learners using this textbook (and the chosen TED Talks) may not develop the necessary academic vocabulary knowledge needed for their studies. This may lead to comprehension difficulties, fixation, or loss of focus, as the learners are unable to process the input they hear in real time (Goh 2000). These results suggest that learners need more exposure to these—and other—academic items in TED Talks; otherwise, they may not be sufficiently prepared to develop their academic vocabulary knowledge as they progress through the textbook.

Second, while general and academic MWUs are evident in the TED Talks, their relative frequency in each unit is not sufficiently covered, recycled, or repeated sufficiently throughout the textbook. Similar to Wang and Liu's (2023) findings, the lack of exposure of MWUs

means that these words may remain unlearned, as their infrequent repetition does not allow for the learners to develop their knowledge of these terms. As O'Loughlin (2012) found in his study, the lack of repetition opportunities may cause vocabulary knowledge difficulties for learners. For example, if learners have not mastered the words from previous units or textbook levels, it may be incorrectly assumed that they have acquired adequate vocabulary knowledge. Wang and Liu (2023) emphasized the important correlation between repeated encounters of collocations and vocabulary knowledge for better vocabulary gains. These results suggest that although coverage of general vocabulary may be sufficient, academic vocabulary and MWUs are not as frequently repeated when developing textbook materials. Unit activities could feature correlating language to link the textbook to the chosen TED Talk. As Keynote 2 is considered a textbook to be used for academic purposes, the need to include both academic and MWU items within the textbook activities is more pertinent, as learners may encounter lexical difficulties if they do not repeatedly meet these items before listening. With over 3500 TED Talks (Ted.com accessed 10 September 2024) currently available, more attention to choosing TED Talks and facilitating textbook activities that include a minimum of 4% academic vocabulary should be prioritized so learners can recycle and acquire these words. From these results, there is a strong suggestion that the textbook activities could include more suitable content to expose learners to repeated vocabulary to support the vocabulary knowledge needed for the chosen TED Talks. As the next section discusses, several possible solutions could provide learners with the necessary exposure to academic vocabulary and MWUs when using Keynote 2.

6. Pedagogical Implications

There are three main implications for simplifying the lexical demands of both the TED Talks and the textbook from this study. First, concerning the repetition of vocabulary, academic or multiword-unit-specific word lists could be used to prime learners before listening or help scaffold their learning while listening with Keynote 2. As the results showed, although the chosen TED Talks included specific word types, the textbook did not use these in activities. Therefore, specific word lists could help facilitate learning by allowing learners to encounter these academic words and MWUs more frequently before they listen so they can focus on their comprehension while listening. Hloba (2016) suggested that learners can use transcripts to identify the words before completing information transfer activities (i.e., gap fills) to prepare learners to focus on form in productive activities (i.e., spoken or written post-listening tasks). Using word lists can help facilitate explicit vocabulary instruction to help learners pay attention to new or unknown items throughout the listening process (Goh 2000).

Next, supplementary materials (such as visual prompts) could help support learners in developing their academic vocabulary knowledge when using Keynote 2. As Table 8 shows, the relative frequency of MWUs appeared more in the chosen TED Talks than in the textbook content. One approach to bridge this gap is to introduce learners to vocabulary strategies (i.e., predicting and inferencing) to help develop their vocabulary knowledge of MWUs. For example, learners could be provided with visual prompts or matching activities to help introduce some of the low-frequency units that are included in the TED Talks. Using materials that help learners to predict or learn the MWUs allows learners to engage in real-time comprehension. Learners can then address their lexical difficulties by predicting meaning, inferencing potential topic and vocabulary knowledge gaps, and connecting images to interpretations of the TED Talks. Simultaneously, as Field (2008) noted, teachers can provide learners with more exposure to real-world topics and real-time listening practices using authentic resources that scaffold learners' vocabulary knowledge for more difficult TED Talks in Keynote 2.

Finally, teachers can consider using other textbook activities and alternative TED Talks for more adequate practice within their teaching contexts. As this study's analysis of Keynote 2 has shown, the vocabulary demands for each unit and the specific listening text can differ from the suggested vocabulary exposure needed in the textbook. More attention

in choosing units and resources that are reflective of the support needed to achieve the specified vocabulary knowledge level could provide more effective listening practices when using the textbook or alternative resources. As previous studies showed, there is a discrepancy between the textbook's vocabulary demands of 2000–4000 words and the TED Talks' vocabulary demands of 3000–5000 words (see Tables 1 and 2). Mo and Bi (2024) suggested two principles to ensure more sufficient exposure to vocabulary in textbooks. First, words found in all the frequency lists should be representative of the vocabulary found in both the TED Talks and the textbook. Second, more sophisticated mid-frequency and low-frequency vocabulary from the TED Talks should be included in textbooks to enhance the vocabulary scope of textbook materials and allow for learners to learn words that they may or may not have previously learned. Therefore, more attention in choosing textbooks or resources that carefully align with these vocabulary demands could provide learners with more ample scaffolding, repetition, and exposure to vocabulary that helps learners to maximize their vocabulary learning in listening practices.

7. Limitations and Future Research

This study has two main limitations. First, the current analysis of TED Talks used in one commercial textbook cannot represent the entirety of this or any other textbook series. Currently, Keynote has six textbooks in the series, ranging between elementary to advanced proficiency levels, which use TED Talks as their primary listening input. Furthermore, Cengage Learning has developed another textbook series (i.e., World English) that also uses TED Talks. Further lexical analysis and comparison of these two titles could help establish if there is any correlation between the lexical demands of the chosen TED Talks and the textbook materials. For future research, a larger study to include the TED Talks used in all the Keynote (i.e., Keynote 1, 3, 4, Advanced, and Professional) and World English series could focus on identifying if the lexical choices are appropriate and contain suitable coverage of academic vocabulary. This analysis could also help to validate if the TED Talks and textbook findings from Keynote 2 in this study are reiterated in the whole Keynote series and other textbooks.

Second, although this study focused on academic vocabulary and MWUs, further research is needed to examine specific word types in more depth. Using the NAWL (Browne et al. 2013) could further validate or create new thresholds considering the established 4% suggested coverage from the AWL (Coxhead 2000). Further, using tools such as the academic spoken word list (ASWL) (Dang et al. 2017) could provide a more thorough corpus analysis of spoken texts to help establish specific spoken word lists for each textbook. The findings from such studies can then assist teachers and learners in comprehending TED Talks used in commercial textbooks. In turn, a pedagogically orientated study could be conducted to investigate if these TED-Talk-specific word lists help learners to successfully develop their vocabulary knowledge. Measuring learners' vocabulary knowledge by analyzing their written notes and summaries for MWUs and academic vocabulary content could also provide valuable insights into the use of word lists and their potential to address learners' individual vocabulary difficulties.

8. Conclusions

This paper presented an in-depth lexical analysis of 12 TED Talks and the unit content used in the commercial textbook, Keynote 2. The results showed that the TED Talk selections and the textbook provided learners with limited repetition activities, limited task practices, and limited exposure to academic vocabulary. Specifically, three units included a higher percentage of mid-frequency and low-frequency words, which were lexically more difficult than the textbook's prescribed level of Pre-intermediate learners. Further, the NAWL and academic vocabulary analysis showed that the TED Talks and the textbook did not include recycling or repetitive opportunities. Similarly, the MWUP analysis showed that the TED Talks' MWU items were not highly prioritized or recycled in textbook activities. Therefore, teachers and learners need to identify the potential difficulty of these word

types in listening texts to prime learners sufficiently for the listening text. Creating supplementary academic and MWU-specific word lists can help prepare learners to identify these terms before they watch the TED Talks. With further development of textbook materials, lexically appropriate listening tasks can be tailored to supplement learners using these textbooks to provide them with the scaffolding needed to achieve their language learning goals.

Author Contributions: Conceptualization: N.M.-R. and S.B.; Methodology: N.M.-R. and S.B.; Software: S.B.; Validation: N.M.-R. and S.B.; Formal analysis: N.M.-R. and S.B.; Investigation: N.M.-R. and S.B.; Resources: N.M.-R.; Data curation: N.M.-R. and S.B.; Writing—original draft preparation: N.M.-R. and S.B.; Writing—review and editing: N.M.-R. and S.B.; Visualization: N.M.-R. and S.B.; Supervision: N.M.-R.; Project administration: N.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ackermann, Kirsten, and Yu-Hua Chen. 2013. Developing the academic collocation list (ACL)—A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12: 235–47. [CrossRef]

Anthony, Laurence. 2023. AntWordProfiler (Version 2.2.1) [Computer Software]. Available online: https://www.laurenceanthony.net/software/antwordprofiler/ (accessed on 10 September 2024).

Ashton, Guy, and Lou Burnard. 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Bauer, Laurie, and Ian Stephen Paul Nation. 1993. Word families. *International Journal of Lexicography* 6: 253–79. [CrossRef]

Benson, Stuart, and Naheen Madarbakus-Ring. 2021. A Comparison of Textbook Vocabulary Load Analysis. Vocabulary Learning and Instruction. Available online: https://vli-journal.org/wp/vli-issue-102/ (accessed on 10 September 2024).

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14: 275–311. [CrossRef]

Bohlke, David, Paul Dummett, Lewis Lansford, and Helen Stephenson. 2016. *Keynote 2*. Boston: Cengage Heinle.

Browne, Charles, Brent Culligan, and Joseph Philips. 2013. The New General Service List Website. Available online: https://www.newgeneralservicelist.com/ (accessed on 10 September 2024).

Byrd, Pat, and Averil Coxhead. 2010. On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL* 5: 31–64.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34: 213–38. [CrossRef]

Coxhead, Averil, and Roz Walls. 2012. TED Talks, Vocabulary and Listening for EAP. *TESOL ANZ* 20: 55–68. Available online: https://www.tesolanz.org.nz/wp-content/uploads/2019/12/TESOLANZ_Journal_Vol20_2012.pdf#page=66 (accessed on 10 September 2024).

Coxhead, Averil, Thi Ngoc Yen Dang, and Shota Mukai. 2017. Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes* 30: 66–78. [CrossRef]

Dang, Thi Ngoc Yen, Averil Coxhead, and Stuart Webb. 2017. The academic spoken word list. *Language Learning* 67: 959–97. [CrossRef]

Davies, Mark. 2008. The Corpus of Contemporary American English: 520 Million Words, 1990–Present. Available online: http://corpus.byu.edu/coca/ (accessed on 10 September 2024).

Eguchi, Masaki. 2021. Multi-Word Units Profiler. (Version 2.0.0) [Computer Software]. Available online: https://multiwordunitsprofiler.pythonanywhere.com (accessed on 10 September 2024).

Elk, Carolyn K. 2014. Beyond Mere Listening Comprehension: Using TED Talks and Metacognitive Activities to Encourage Awareness of Errors. *International Journal of Innovation in English Language Teaching and Research* 3: 215–47. Available online: https://search.proquest.com/docview/1655287119?accountid=14782 (accessed on 10 September 2024).

Erman, Britt, and Beatrice Warren. 2000. The idiom principle and the open-choice principle. *Text* 20: 29–62. [CrossRef]

Field, John. 2008. *Listening in the Language Classroom*. New York: Cambridge University Press.

Foster, Pauline. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Edited by Martin Bygate, Peter Skehan and Merill Swain. Harlow: Longman, pp. 75–93.

Gardner, Dee, and Mark Davies. 2014. A new academic vocabulary list. *Applied linguistics* 35: 305–27. [CrossRef]

Goh, Christine. 2000. A cognitive perspective on language learners' listening comprehension problems. *System* 28: 55–75. [CrossRef]
Hajiyeva, Konul. 2015. A corpus-based lexical analysis of subject-specific university textbooks for English majors. *Ampersand* 2: 136–44. [CrossRef]
Heatley, Alex, Ian Stephen Paul Nation, and Averil Coxhead. 2002. Range Computer Programme. Available online: https://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed on 10 September 2024).
Hloba, Olena. 2016. *TED Speeches as a Tool to Improve Listening Skills of Students Majoring in Psychology*. Rome: Edizioni Magi, vol. 4, pp. 1–4. Available online: https://enpuir.npu.edu.ua/handle/123456789/10569 (accessed on 10 September 2024).
Laufer, Batia. 1992. How much lexis is necessary for reading comprehension? In *Vocabulary and Applied Linguistics*. Berlin and Heidelberg: Springer, pp. 126–32.
Liu, Chen Yu, and Howard Hao-Jen Chen. 2019. Academic spoken vocabulary in TED Talks: Implications for academic listening. *English Teaching and Learning* 43: 353–68. [CrossRef]
Macalister, John, and Ian Stephen Paul Nation. 2020. *Language Curriculum Design*. London: Routledge.
Martinez, Ron, and Norbert Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33: 299–320. [CrossRef]
Matsuoka, Warren, and David Hirsh. 2010. Vocabulary Learning through Reading: Does an ELT Course Book Provide Good Opportunities? *Reading in a Foreign Language* 22: 56–70.
McCarthy, Michael, and Ronald Carter. 2006. This that and the other: Multi-word clusters in spoken English and visible patterns of interaction. In *Explorations in Corpus Linguistics*. Edited by Michael McCarthy. Cambridge: Cambridge University Press, pp. 7–26.
Mo, Junhua, and Peng Bi. 2024. Evaluation of Vocabulary Use in EFL Textbooks: Evidence from Curriculum Words. *English Teaching and Learning* 1–16. [CrossRef]
Mojgan, Rashtchi, and Mohammad Mazraehno Reza Tollabi. 2019. Exploring Iranian EFL learners' listening skills via TED Talks: Does medium make a difference? *Journal of Language and Education* 5: 81–97. [CrossRef]
Morrow, Keith. 1977. Authentic texts and ESP. In *English for Specific Purposes*. Edited by Susan Holden. London: Modern English Publications, pp. 13–17.
Nation, Ian Stephen Paul. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review* 63: 59–82. [CrossRef]
Nation, Ian Stephen Paul. 2012. The BNC/COCA Word Family Lists (17 September 2012). Unpublished Paper. Available online: www.victoria.ac.nz/lals/about/staff/paul-nation (accessed on 10 September 2024).
Nation, Ian Stephen Paul. 2022. *Learning Vocabulary in Another Language*, 3rd ed. Cambridge: Cambridge University Press.
Nation, Ian Stephen Paul, and Stuart A. Webb. 2011. *Researching and Analyzing Vocabulary*. Boston: Heinle, Cengage Learning.
Nation, Ian Stephen Paul, Dongwang Shin, and Lynn Grant. 2016. Multiword units. In *Making and Using Word Lists for Language Learning and Testing*. Edited by Ian Stephen Paul Nation. Amsterdam: John Benjamins.
Nurmukhamedov, Ulugbek. 2017. Lexical coverage of TED Talks: Implications for vocabulary instruction. *TESOL Journal* 8: 768–90. [CrossRef]
Nurmukhamedov, Ulugbek, and Randall Sadler. 2011. Podcasts in four categories: Applications to language learning. In *Academic Podcasting and Mobile Assisted Language Learning*. Edited by Betty Rose Facer and M'hammed Abdous. Portland: Book News, pp. 176–95.
O'Loughlin, Richard. 2012. Tuning in to vocabulary frequency in coursebooks. *RELC Journal* 43: 255–69. [CrossRef]
Romanelli, Frank, Jeff Cain, and Patrick J. McNamara. 2014. Should TED talks be teaching us something? *American Journal of Pharmaceutical Education* 78: 113. [CrossRef]
Schmitt, Norbett, and Diane Schmitt. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching* 47: 484–503. [CrossRef]
Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31: 487–512. [CrossRef]
Siyanova, Anna, and Norbett Schmitt. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review* 64: 429–58. [CrossRef]
Siyanova-Chanturia, Anna, and Ron Martinez. 2015. The idiom principle revisited. *Applied Linguistics* 36: 549–69. [CrossRef]
Sun, Ye, and Thi Ngoc Yen Dang. 2020. Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System* 93: 102279. [CrossRef]
TED Talks. 2024. TED Talks Website. Available online: https://www.ted.com/talks (accessed on 10 September 2024).
Van-Zeeland, Hilde, and Norbett Schmitt. 2013. Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics* 34: 457–79. [CrossRef]
Wang, Chen, and Yuhua Liu. 2023. Measuring Vocabulary Use in Chinese Tertiary Textbooks: Potentials for Incidental Vocabulary Learning. *Asian Journal of English Language Teaching* 32: 131–54.
Warnby, Marcus. 2023. Academic vocabulary knowledge among adolescents in university preparatory programmes. *Journal of English for Academic Purposes* 61: 101203. [CrossRef]
West, Michael. 1953. *A General Service List of English Words: With Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. Harlow: Longman.
Wingrove, Peter. 2022. Academic lexical coverage in TED talks and academic lectures. *English for Specific Purposes* 65: 79–94. [CrossRef]

Wood, David. 2002. Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal* 20: 1–15. [CrossRef]
Wood, David C., and Randy Appel. 2014. Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes* 15: 1–13. [CrossRef]
Yang, Lu, and Averil Coxhead. 2020. A corpus-based study of vocabulary in the new concept English textbook series. *RELC Journal* 53: 1–15. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

The Influence of the L1 on L2 Collocation Processing in Tamil-English Bilingual Children

Roopa Leonard [1,*], Holly Joseph [2] and Michael Daller [1]

[1] Department of English Language and Applied Linguistics, University of Reading, Reading RG6 6UR, UK; michael.daller@yahoo.com
[2] Institute of Education, University of Reading, Reading RG6 6UR, UK; h.joseph@reading.ac.uk
* Correspondence: roopa.leonard@gmail.com

Abstract: This study examines the influence of Tamil (L1) on the processing of English (L2) collocations during reading for Tamil-English bilingual children. Building on existing research in formulaic language, we used an online processing tool to investigate whether cross-linguistic transfer can be extended beyond single lexical items to collocations in bilingual children, a population that is under-represented in this research area. Fifty-eight children aged 9–10 years from a school in Chennai, India, took part. Using self-paced reading, children's reading times were measured for both congruent (with equivalent in L2) and incongruent (without equivalent in L2) English collocations embedded in short passages. There were two reading modes (single and chunk), which allowed reading times for the whole collocations and the individual words of the collocations to be examined. Results showed that children read congruent collocations more quickly than incongruent collocations in both modes. For congruent collocations, children read the second word more quickly than the first word, but the reverse was true for incongruent collocations. These results suggest that the L1 (Tamil) is activated during the processing stage of reading English collocations for Tamil-English bilingual children in this context.

Keywords: collocations; formulaic language; cross-linguistic influence; L2 vocabulary

1. Introduction

With an increasing number of children learning English as a second or additional language in India, the UK, and around the world, it is important to develop our understanding of how second language (L2) vocabulary learning in children takes place, especially in relation to the influence of the first language (L1) on this process. Understanding this can help educators support second language learners by effectively using L1 to teach L2 vocabulary (Zulfikar 2019). As mentioned by Hiebert et al. (2019), research in the area of L1 influence on L2 vocabulary learning has until recently focused on single words. We often think of vocabulary in terms of individual words, but words actually co-occur frequently in systematic ways to form collocations and other multiword units, that is, formulaic language. Since we know that vocabulary is also acquired and stored in multiword units (Schmitt 2010), not just as single words, it is essential that we expand our knowledge of the role of the L1 in how bilingual children process formulaic language in the L2. Studying the processing of formulaic language can further our understanding of how monolingual and bilingual lexicons are formed and how they differ. This knowledge can help in the design of vocabulary instruction material for children as well as inform vocabulary teaching practices in the classroom (Lindsay and Gaskell 2010).

Collocations are a subset of formulaic language that can be broadly defined as words that occur together more frequently than would be expected by chance (Carrol and Conklin 2020), e.g., strong wind. A simple but rather broad definition of a collocation is a frequently recurring two-to-three word syntagmatic unit (Henriksen 2013). Hunston (2002) describes

it as the tendency of two words to co-occur or the tendency of one word to attract another. Wray (2002) observes a critical difference between collocations and other kinds of formulaic language such as idioms: collocations are more "fluid" in nature, whereas idioms are fixed, e.g., *to make a mountain out of a molehill* is a fixed expression, while though *rain* often collocates with *heavy*, it is still associated with other words. Studies have shown that collocations, like other kinds of formulaic language, are processed more quickly than nonformulaic language (Bonk and Healy 2005; Siyanova-Chanturia and Sidtis 2018; Wolter and Gyllstad 2011). A number of studies have investigated how the correspondence between L1 and L2 contributes to online processing of formulaic language (Carrol et al. 2016): if formulaic language shares form and meaning across both a speaker's languages, then speakers perform better on comprehension and production tasks, which could be indicative of faster processing (Paquot and Naets 2015; Pritchett et al. 2016).

1.1. Cross-Linguistic Processing of Collocations

Although older models of bilingual lexical storage and access, such as the Revised Hierarchical Model (RHM) (Kroll and Stewart 1994), viewed the L1 and L2 lexica as two separate entities, the more recent Bilingual Activation Model (BIA) (Dijkstra and Van Heuven 1998) and Multilink Model (Dijkstra and Rekké 2010), suggest there is one integrated lexicon in which both L1 and L2 lexical items are stored. Jiang's model of L2 lexical representation and development (Jiang 2000) posits that the acquisition of L1 words with an L2 equivalent is quicker because the learner already has access to the required semantic and syntactic information at the lemma level and thus only needs to acquire the phonological and orthographic information at the lexeme level. However, when a word does not have an L2 equivalent, the learner has to acquire the L2 information at both the lexeme and lemma level, and this process requires more effort and time.

Collocations can vary considerably from language to language (Wolter and Gyllstad 2011), and it has been observed that this variation is arbitrary (Henriksen 2013), i.e., there is often no logical or grammatical explanation for why certain words collocate with each other and others do not. The influence of the L1 on the acquisition of collocations (intralexical links) is an area that has attracted a great deal of scholarly attention over the last decade, as detailed below. Although there are different explanations for the influence of L1 on L2 collocation acquisition, the understanding of how this influence works is still in its early stages. This influence of L1 has mostly been examined by studying the differences in the production, reception, and processing of congruent collocations (collocations that have a direct equivalent in the learners' native language) and incongruent collocations (collocations that do not have a direct equivalent in the learners' native language).

Yamashita and Jiang (2010) presented verb-noun and adjective-noun collocations in an acceptability judgment task to Japanese EFL (English as a Foreign Language) and ESL learners. The researchers found that while there was no difference in processing times or error rates between congruent and incongruent collocations for the monolingual speakers (as expected), for both ESL (English as a Second Language) and EFL Japanese learners, the error rates were higher for the incongruent phrases. Additionally, the EFL learners had slower reaction times for the incongruent phrases, but this was not the case for the ESL learners. They also used a cloze test with the Japanese speakers to determine the difference in levels of L2 proficiency between the EFL and ESL learners—there was a significant difference in mean scores between both groups, with the ESL learners outperforming the EFL learners. This led the researchers to conclude that L2 collocations are processed independently of the L1 lexicon only at the later stages of language acquisition, i.e., when the learners become more proficient. This means that the influence of the L1 is greater during the initial stages of language acquisition and gradually subsides in the later stages.

Wolter and Gyllstad (2011, 2013) investigated collocational priming on congruent and incongruent collocations in monolingual speakers of English and Swedish learners of English and found similar results to Yamashita and Jiang's study. For the Swedish learners of English, the processing times for the target word for the incongruent collocations were

significantly longer than the same measure for the congruent collocations. Wolter and Gyllstad concluded that the advantage for congruent collocations may be due to a lexical priming effect: the knowledge of the collocation in the L1 primes their knowledge of the equivalent L2 collocation, i.e., the congruent collocation, thus reducing the processing time for such collocations. Thus, the L1 appears to be providing easier access to L2 collocations that have an equivalent in the L1, which is not possible for L2-only collocations. In terms of frequency, for both monolingual speakers and the Swedish learners of English, the biggest predictor of reaction times was the frequency of the collocations in relation to the English corpus. Wolter and Gyllstad cite other studies (e.g., Durrant and Schmitt 2010), which also found that high-proficiency language learners are sensitive to frequency effects just like monolingual speakers. The findings from this second study by Wolter and Gyllstad (2013) suggest that advanced L2 learners are sensitive to frequency effects not only at the word level, but also at the collocational level. Additionally, it must be noted that collocational frequency was the biggest predictor of reaction times, not word frequency. This indicates that both monolingual speakers and the Swedish learners of English processed the collocations holistically as single units and not as separate words.

Wolter and Yamashita (2018) followed up on these previous studies by conducting a study that examined the effects of word frequency, collocational frequency, L1–L2 congruency, and language proficiency on L2 collocational processing in Japanese speakers of English (intermediate and advanced) and monolingual English speakers. With respect to frequency, the results showed that the monolingual English speakers and the advanced Japanese speakers of English showed a greater sensitivity to collocational frequency than the intermediate Japanese speakers of English. Interestingly, further analysis of the word frequency and collocational frequency with both groups of Japanese learners showed that, with increased proficiency there was a shift away from reliance on word frequency to reliance on collocational frequency. This indicates that with increasing proficiency, L2 learners move towards monolingual-like collocational processing. However, it must be noted that even the most advanced L2 Japanese learners in this study relied more heavily on word frequency than the monolingual English speakers, which supports the conclusions of previous studies that even advanced L2 learners have difficulties with acquiring and processing L2 collocations.

Taking the results of these studies together, it is clear that across different contexts and with different groups of learners, the L1 influences L2 collocational processing, i.e., there is a congruency effect in which the L1 is activated during L2 collocational processing. Congruent collocations are recognized and processed more accurately than incongruent collocations, i.e., L1-only or L2-only collocations. Also, once a collocation is registered in the L2 lexicon, its L2 frequency influences its processing.

1.2. Sequential Bilingualism in Children

In early second language acquisition (ESLA), children encounter a second language after beginning to acquire their first, typically when they start childcare, nursery, or school. This process, known as sequential (successive) bilingualism, contrasts with simultaneous bilingualism, where children hear both languages at birth. Unlike the typical view of L2 acquisition in adults, where a second language is introduced after the first is fully developed, children who are sequential bilinguals usually hear their first language (L1) at home and are exposed to the second language (L2) in educational settings while there L1 is still developing.

McLaughlin (1995, 2013) highlights several factors that influence these components and processes, such as cultural, linguistic, and social differences, along with variations in educational systems, individual attitudes, and cognitive abilities. Paradis et al. (2011), studying 169 sequential bilinguals aged four to seven, found that internal factors like age, nonverbal intelligence, and phonological short-term memory predicted L2 proficiency better than external factors like length of exposure or the richness of English use at home. A study on Russian-Hebrew and English-Hebrew bilingual children (Armon-Lotem et al.

2014) revealed that L2 proficiency was closely tied to attitudes towards the heritage language and L2, with greater proficiency in communities with mixed language use, such as the Russian-Hebrew community.

This brief overview of sequential bilingualism in children shows that there are many different variables that influence the development of both languages in a sequentially bilingual child, which differ from a bilingual adult. The children in this study are sequential bilinguals, and it is important to consider these factors when analyzing how Tamil influences their processing of English collocations.

1.3. Cross-Linguistic Influence on Language Processing in Bilingual Children

It is worth noting that most of the previous research examining the extent of L1 influence on L2 collocational processing has been conducted with adults (e.g., Yamashita and Jiang 2010; Wolter and Gyllstad 2011, 2013; Wolter and Yamashita 2018). Since differences have been observed between cross-linguistic activation in bilingual children and adults (Paradis et al. 2011; Yip and Matthews 2000), it is important to establish whether the pattern of effects observed in adults is also observed in children. From studies that have investigated the cross-linguistic influence in children and adults (Murphy 2014; Paradis et al. 2011), it seems that bilingual children are more prone than adults to quantitative cross-linguistic influence, i.e., they are more likely to use their knowledge of one language while producing constructions in their other language. A recent case study by Babatsouli and Nicoladis (2019) investigated cross-linguistic influence in a 4-year-old bilingual child in the context of fixed expressions and found that the child showed evidence of cross-linguistic transfer in her use of collocations. However, whether older bilingual children exhibit this pattern too is as yet under-researched.

1.4. Socioenvironmental Background of the Study

In the highly multilingual setting of India, English has the constitutional status as one of the two official languages and serves as a lingua franca (Ayyar 1993), particularly in large cities such as Chennai. In Chennai, English-medium schools include: (1) schools that follow an international curriculum that cater to high-income families; (2) elite private schools that follow the state curriculum (Tamil Nadu State Board) or a national curriculum (Central Board of Secondary Education); and (3) low-income state and private schools that cater to children from less privileged families. Schools in this last category are only English-medium in a formal sense—although the whole curriculum is meant to be taught in English, classroom interactions are mostly in the regional language since teachers themselves struggle with communicating effectively in English. In most of these schools, class sizes range from 40 to 60, and virtually all teachers of English have learned English in the same education system (Ponnuchamy 2012). Typically, children who attend these low-income schools have very minimal or no contact with English outside the school setting. As noted by Ponnuchamy (2012), the state government's efforts to promote English at these schools have been met with criticism due to poor school facilities, a lack of adequate and effective training for teachers, and teachers' inefficiency at communicating in English. Gargesh (2006) observes that after approximately 12 years of education in state or low-income private schools (English-medium), most of these students in Tamil Nadu are unable to cope with the academic demands of tertiary education in colleges and universities, which all have English as the medium of instruction. The school that participated in this study belongs to this category and caters to children from low-income families in Chennai. Since these children are not very proficient in English and their main source of English exposure is from an L2 speaker who is not very proficient (their teacher), it is likely that these children's processing of English collocations will be influenced by their Tamil. This lack of English competence of English teachers world-wide is an issue that has been well documented (Dearden 2014).

1.5. Rationale for the Current Study

The current study examined L1 influence on L2 collocational processing in bilingual English-Tamil speaking children. While previous research has focused mostly on collocation processing in adults (e.g., Yamashita and Jiang 2010; Wolter and Gyllstad 2011, 2013; Wolter and Yamashita 2018), the current study focuses on collocation processing in children, who we know differ from adults in important ways (Murphy 2014; Paradis et al. 2011), which is likely to affect the extent to which their L1 impacts their L2 processing. By examining transfer between languages that are orthographically different, we were able to broaden our understanding of L1 activation during L2 processing beyond the usual suspects (European alphabetic languages). Finally, we explored the role of proficiency and vocabulary in both languages, as previous research suggests that with advanced proficiency in L2, L1 influence of formulaic language diminishes (Carrol et al. 2016).

This study used self-paced reading to examine whether Tamil-English bilingual children showed a congruency effect when reading passages containing congruent and incongruent collocations. We had two main research questions:

1. Is the L1 activated when Tamil-English bilingual children process collocations in L2?
2. What is the relationship between the proficiency levels/vocabulary size of children and the time they take process congruent versus incongruent collocations?

In relation to the first research question, it was predicted that congruent collocations would be processed more quickly than the incongruent ones, which would be an indication that Tamil (L1) was activated during the processing of English (L2) collocations. This prediction followed previous studies showing that the influence of the L1 is greater at initial stages of language acquisition (e.g., Yamashita and Jiang 2010). Furthermore, it was predicted that congruency would have a priming effect (when the first word of the collocation prepares the reader for the second word) on the second word of collocations during the processing stage of collocations for Tamil-English bilingual children (Wolter and Gyllstad 2011).

In relation to the second research question, it was predicted that children with a larger vocabulary in Tamil than in English would show a larger congruency effect than those with a larger vocabulary size in English. Additionally, it was predicted that children with higher levels of English proficiency would show a smaller congruency effect than those with lower proficiency. This would be consistent with previous empirical studies (Bonk and Healy 2005; Wolter and Gyllstad 2011).

2. Materials and Methods

2.1. Participants

Fifty-eight 5th grade students (34 girls) participated in this study. All participants were 9–10 years old and attended an English-medium primary school in Chennai, India. The selection of this age group was based on the requirement of a certain level of reading ability and English proficiency for the self-paced reading task. All participants had Tamil as their L1 and English as their L2, and they did not speak any other languages. The medium of instruction in their school was English, but they also learned Tamil as a separate subject. The students had 6–7 years of English-medium education since the beginning of their schooling. The selected school was a low-tier private school with affordable fees and catered to children from low-income families who desired English-medium education for their children but could not afford more expensive, well-resourced private schools. Although English was the medium of instruction in their school, the students had limited exposure to English outside school, and Tamil was their preferred language of communication. However, they had more exposure to written English than written Tamil due to the English language textbooks used in their school.

2.2. Preparation of Stimuli

Since the definition of collocations is inclusive of various kinds of word combinations, a critical part of every collocation study is determining and defining the kind of collocations

to include in the study. In this study, collocations were extracted from the textbooks of the participating children instead of corpora. This decision was made because the children had limited exposure to English outside the classroom, and using a corpus might not reflect their actual exposure to English. Additionally, using the children's textbooks ensured that the selected collocations were already familiar to the learners. This approach aligns with the findings of Northbrook and Conklin (2019), who observed that beginner Japanese learners of English were more sensitive to formulaic language extracted from their learning materials rather than from a corpus based on frequency. The frequency of the collocations was determined using the BNC. The Oxford Collocation Dictionary and Collins CoBuild English Dictionary were consulted to ensure that the selected collocations were used in standard English. The collocations were chosen based on their frequency (from the 5000 most frequent English words according to the BNC) and an MI (Mutual Information) score above 3, which indicates a recognizable association between collocates (Hunston 2002). This shortlist of collocations was crosschecked with the children's textbooks as mentioned above, and then the final list of collocations was chosen. The selected collocations were then translated into Tamil by the researcher and cross-checked for accuracy and congruency by two other adult first language speakers of Tamil who are also proficient in English. An equal number of incongruent and congruent collocations were chosen based on their fit into the mini stories.

The stimuli for the experiment consisted of eleven mini stories, each comprising three to four sentences. Each story contained a mixture of four to five congruent and incongruent collocations, with a maximum of two collocations per sentence. The collocations consisted of adjective+noun, verb+noun, and verb+adverb combinations. The length and number of collocations in the stories were controlled to ensure comparability between single and chunk modes of presentation. An example of a mini story with collocations in bold was provided.

Example mini story (with collocations in bold):

> It was the **rainy season** (incongruent), and Timmy could not go out to play. It was the **first time** (congruent) this year that he had seen such a **strong wind** (congruent) blowing outside. In **broad daylight** (incongruent), he saw all the birds **fly away** (incongruent) to take shelter in the trees.

2.3. Design

The study followed a 2 × 2 within-participants design with congruency and presentation mode as independent variables. Congruent collocations could be translated word-for-word from English to Tamil while retaining the same meaning, while incongruent collocations lacked equivalents in Tamil. The presentation mode consisted of single and chunk modes. In the single mode, sentences were presented word-by-word, requiring participants to press a button to proceed to the next word, including the collocations. In the chunk mode, each sentence was presented in two/three-word units (chunks), allowing participants to read each collocation as a single unit. Care was taken to ensure the collocations remained within the chunk, and so certain units contained three words (usually a conjunction or article) to accommodate this. This was done so that the reading times for collocations as a whole could be recorded, as well as each individual word to investigate whether there was a priming effect (when the first word of the collocation prepares the reader for the second word). This was how the collocations were presented, but on later reflection, we decided that this manipulation is not relevant for most of the analyses. We will report minimally on these conditions for transparency but not discuss them in detail. The first five stories were presented in single mode, and the remaining six were presented in chunk mode for all participants. Thus, each participant was exposed to four experimental reading conditions: single congruent, single incongruent, chunk congruent, and chunk incongruent. The dependent variable was reading time, measured in milliseconds, with vocabulary and proficiency measures in both languages serving as covariates.

2.4. Research Instruments

2.4.1. X-lex Tests

The X-lex 5000 test, developed by Meara and Milton (2003), was used to measure receptive vocabulary size. It includes both real words and pseudowords and focuses on vocabulary breadth. The test employs a Yes/No format, where participants mark the words they know. The test has been found reliable and valid for screening, placement, and measuring average vocabulary size. Although it was originally designed for EFL students at the college level, it has successfully been used with children (Milton 2006; Daller and Ongun 2018). It uses the 5000 most frequent English words in the BNC. Each test consists of 100 words, with 20 words from each of the first five frequency bands, and includes 20 pseudowords that are phonologically and orthographically valid but are not actual words, e.g., flimsale. For every correct word, the participants received 50 points, and they lost 250 points for every pseudoword they marked as known. Thus, the pseudowords act as a correction formula to control for response bias and cheating.

The X-lex has been reported as reliable and valid for screening and placement purposes and also as a measure of average vocabulary size (Milton 2007). Due to the fact that it is easy to administer and score and also allows for a sampling of a large number of items, it has been used widely in vocabulary testing in second language research (Harsch and Hartig 2015). In a study that measured the relationship between X-lex scores and reading and listening skills, Harsch and Hartig (2015) found a significant positive correlation between the X-lex scores and the reading proficiency levels of the learners.

For the X-lex English test in this study, the existing test was used (Meara and Milton 2003). For the X-lex Tamil test, the researcher developed the test using the Tamil corpora available on the Sketch Engine website (https://www.sketchengine.eu/) (accessed on 23 September 2024) and in accordance with the pattern of the English X-lex test. The test was piloted with seven Tamil-English bilingual children and three Tamil-English bilingual adults, and no changes were needed.

2.4.2. C-Test

The C-test, an integrative written test of general language proficiency, was administered to measure overall English language skills. It consists of authentic short texts covering different topics, where half of every second word is deleted, excluding one-letter words, proper nouns, and numbers. Test-takers fill in the blanks. In most cases, there is only one right answer for each blank. Each right answer gets one point, and there is no negative marking.

Since its introduction, various studies have been carried out to validate the reliability of the C-test as a test of general language proficiency and have reported positive results (e.g., Babaii and Ansary 2001; Dörnyei and Katona 1992). In a study designed to test the predictive power of C-test scores on reading ability, Harsch and Hartig (2015) found that the C-test had high correlations with both reading and listening skills. Eckes and Grotjahn (2006) and Grotjahn (2022) report the C-test to be easy to administer, objective, and a reliable measure of language proficiency.

The C-test for this study was developed from scratch based on topics from the children's English textbooks, and there were 50 blanks in total. The test was piloted with Tamil-English children from the relevant age group, and following this, necessary changes were made.

2.4.3. Self-Paced Reading Experiment

Since it gives a measure of processing in real time, self-paced reading has been used in a range of linguistic processing studies—from studying the effects of language switching on reading comprehension (Bultena et al. 2015) to examining the processing of past tense morphology (Pliatsikas and Marinis 2013). There are also numerous studies that have used the self-paced reading task to study the processing of formulaic sequences: Kim and Kim (2012) used it to study the effects of frequency on multiword processing in L2 learners and first language speakers, Schmitt and Underwood (2004) used it to analyze the processing

of component words in formulaic sequences, and Siyanova and Schmitt (2008) used it to study the processing of L2 collocations.

The self-paced reading task was conducted using PsychoPy (Peirce 2007) v1.8, an open-source application designed to control stimulus display and measure timing. It presents stimuli on a computer monitor, and participants press the spacebar to proceed to the next unit. The application records the time between button presses as the reading time. The stimuli, consisting of mini stories, were entered word-by-word for the single mode and in two-word chunks for the chunk mode. The order of presentation was randomized for each participant within the single and chunk blocks. PsychoPy automatically recorded the reading times for each word and chunk in separate files for each participant.

2.5. Procedure

The X-lex tests and C-test were group-administered to the participants in their classrooms (X-lex in the morning and C-test in the afternoon). The participants marked the words they understood and knew how to use in the X-lex tests, without being informed of the presence of pseudowords. The format of the C-test was explained, and participants were given an example to understand the task.

The self-paced reading experiment took place in individual sessions using a laptop in a designated room at the school. The researcher demonstrated the task using an example mini story, followed by a practice trial. Subsequently, participants read the 11 mini stories in a different random order within the single and chunk modes. Each session lasted 30–40 min.

3. Results

Although PsychoPy recorded the reading times for each word and unit during the self-paced reading, only the reading times for the collocations were required for data analysis. The researcher extracted the reading times for the whole collocation in the chunk mode and the reading times for each word of the collocations in the single mode. Since the design was not fully crossed (all stories appeared either in single or chunk mode, not in both), two-way ANOVAs rather than linear mixed effects models were used to analyze the data.

Mean reading times for each condition are shown in Table 1. Note that the reading times on Word 1 and Word 2 were summed in the single condition, to produce overall means for the collocation. As is usually the case with reading time data, the data were not normally distributed, i.e., $p < 0.05$ in the Shapiro–Wilk's test. Based on accepted practice in the field, any reading times that were 2.5 deviations away from the mean for each of the four conditions were excluded (Conklin and Pellicer-Sánchez 2016). This set of data was normally distributed for all conditions ($p > 0.05$ in the Shapiro–Wilk's test), and the analyses were performed on these data. The slower reading time of the single mode can be accounted for by the required key press between words in this mode (not required for chunk mode). Means and standard deviations for vocabulary test scores (out of 5000) and English proficiency test scores (out of 50) as well as Cronbach's alpha for all three tests are listed in Table 2.

Table 1. Mean reading times for congruent and incongruent, and single and chunked collocations (ms). Standard deviations are in parentheses.

	Congruent	**Incongruent**
Single (Word 1 + Word 2)	2322 (807)	2631 (1204)
Chunked	1535 (443)	1742 (470)

Table 2. Means and standard deviations for vocabulary test scores (out of 5000) and English proficiency test scores (out of 50). Cronbach's alpha for all three tests.

	Mean	SD	Cronbach's Alpha
Vocabulary English	2742	957	0.719
Vocabulary Tamil	2161	857	0.761
Proficiency English	32.39	7.54	0.802

3.1. Effects of Congruency, Presentation Mode, and Test Scores

A two-way repeated measures ANCOVA was run to determine the effects of congruency on mean reading times while also accounting for participants' vocabulary knowledge in English and Tamil and for their proficiency in English. The distribution of these scores is seen in Figure 1. There was a main effect of congruency with longer reading times on the incongruent than congruent collocations $F1(1,48) = 50.93$, $p < 0.001$; $F2(1,42) = 16.36$, $p = 0.016$. There was a main effect of presentation mode with longer reading times for single mode than for chunk mode $F1(1,48) = 48.21$, $p = 0.034$; $F2(1,42) = 20.31$, $p = 0.008$. The interactions between congruency and presentation mode were not significant $F1(1,48) = 16.48$, $p = 0.101$; $F2(1,42) = 12.78$, $p = 0.119$.

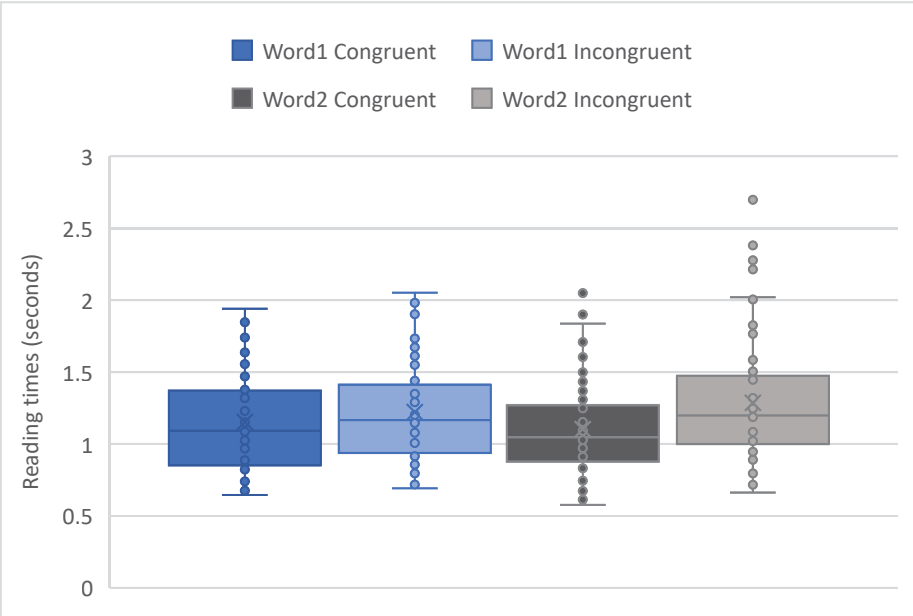

Figure 1. Reading times (in ms) for Word 1 and Word 2 in the congruent and incongruent conditions in single mode.

In terms of the vocabulary and proficiency scores, there were no significant covariate effects (all Fs < 1; all ps > 0.1) on reading times, that is, vocabulary knowledge in English and Tamil as well as English proficiency did not affect how long the children took to read the collocations. Furthermore, there were no significant interactions between congruency and X-lex English Test score, $F1(1) = 0.029$, $p = 0.866$; X-lex Tamil Test score $F1(1) = 0.666$, $p = 0.419$; or C Test score, $F1(1) = 0.122$, $p = 0.729$.

3.2. Congruency Effect on Word 2 in the Single Mode

To further examine the congruency effect on Word 2, Word 2 reading times were isolated and analyzed. Since the congruent and incongruent collocations were not matched for word length and frequency, reading times were transformed into z-scores in order to ensure that reading times for both kinds of collocations could be compared. The difference between the standardized reading times for Word 2 of the incongruent collocations and the standardized reading times for Word 2 of the congruent collocations was calculated

as a new variable: Word 2 Difference. Preliminary analyses showed the values to be normally distributed as assessed by Shapiro–Wilk's test ($p > 0.05$). A Pearson's correlation test was run to examine the correlation between the Tamil and English vocabulary scores and Word 2 Difference. This was performed to see if the vocabulary scores in either language were correlated with the size of the congruency effect on Word 2. The correlation between Tamil vocabulary and Word 2 Difference as well as the correlation between English vocabulary and Tamil vocabulary, were not significant. There was, however, a significant negative correlation between English vocabulary scores and Word 2 Difference $r(58) = -0.84$, $p < 0.001$, showing that children with higher English scores showed a smaller congruency effect on Word 2.

4. Discussion

This study examines L1 influence on L2 processing of collocations in bilingual Tamil-English children. While we know quite a lot about cross-linguistic influence in adults, we know much less about multiword phrases and in younger readers who differ from adult learners in significant ways (see Section 1.2). Results of the self-paced reading experiment supported the congruency effect was on Word 2, as expected (see Wolter and Gyllstad 2011). The lack of positive correlations between the collocation reading times and vocabulary and proficiency test scores was unexpected and will be explored in more detail.

4.1. Congruency

Children read congruent collocations faster than they read incongruent ones even when vocabulary knowledge in English and Tamil, and proficiency in English were controlled for. This robust congruency effect indicates that children activated their L1 (Tamil) while reading L2 (English) collocations, in both presentation modes, i.e., when they were presented with the collocations as single words as well as when they were presented as complete multiword phrases.

From a theoretical standpoint, it is clear that nonselective lexical activation in the Multilink model (Dijkstra and Rekké 2010) and BIA model (Dijkstra and Van Heuven 1998), wherein both the L1 and L2 are activated even though the learner is presented with input only in the L2, is not limited to single words but can also be extended to collocations (Bonk and Healy 2005; Wolter and Gyllstad 2011). It can be assumed that the L1 provides quicker access to L2 collocations that have an L1 equivalent than those L2 collocations that have no L1 equivalent. It is clear that L1 is not suppressed when L2 collocations are activated; instead, there is a process of nonselective activation. It is plausible that when the child encounters the first word of the collocation, the L1 translation is activated along with the L1 collocates, and these L1 collocates activate the L2 collocates, thus allowing for quicker processing times. Based on the congruency effect found in this study, it can be assumed that collocational knowledge is part of the L1 semantic and syntactic information that the learner retains and uses while acquiring L2 vocabulary.

Earlier studies by Yamashita and Jiang (2010); Wolter and Gyllstad (2011), and Siyanova and Schmitt (2008) found that with an increase in L2 proficiency, the L1 influence on collocation acquisition decreases. Although the participants in this study have all been learning English for 5–7 years and also study in a school where the medium of instruction is English, Tamil is dominant in their lives, and this may be why the congruency effect is evident. As seen in the results, there was a lack of positive correlations between English proficiency and the congruency effect. These findings suggest that the relationship between proficiency and congruency effects seen in adults is not necessarily the same for children and support previous research showing this relationship is not always linear (Ding and Reynolds 2019; Fang and Zhang 2021).

4.2. L2 Vocabulary and Congruency Effects

Analyses of Word 2 of the collocation showed that children with larger English vocabularies showed a smaller congruency effect, i.e., the Word 2 Difference was smaller for

those with larger English vocabularies. This is supported by Kroll and Stewart's (1994) theory that the frequency with which bilinguals need to access the L1 translation while processing words in the L2 presumably decreases as they advance in L2 proficiency. This is also supported by a study by Yamashita and Jiang (2010), who found that only the lower proficiency learners showed significantly lower processing times for incongruent collocations. This led the researchers to conclude that L2 collocations are gradually processed independently of the L1 lexicon only at the later stages of language acquisition, i.e., when the learners become more proficient.

Thus, children with larger English vocabularies would not access their knowledge of L1 (Tamil) collocations while reading L2 (English) collocations as frequently as those children with smaller English vocabularies, and this accounts for the smaller congruency effect in the former group.

4.3. Vocabulary and Proficiency Tests and Collocation Reading Times

There was a positive correlation between the English vocabulary scores and Tamil vocabulary scores, as well as between the English vocabulary scores and the English proficiency scores. The test scores from both the vocabulary tests, English and Tamil, did not, however, correlate with the reading times. For the Tamil X-lex test, the words were taken from a corpora of the 5000 most frequently occurring words in written Tamil from the website Sketch Engine, since there is no comparable corpora for spoken Tamil. While the frequency discrepancy between the most frequent words in Tamil and the most frequent words in a child's lexicon could also explain the lack of correlation between the test scores and the reading times, for the Tamil test there is an additional factor that should be taken into consideration. Tamil is diglossic in nature—there is a wide gap between the spoken and written forms of the language. The children in this study are more familiar with the spoken form of Tamil because they encounter it far more in daily life than they encounter the written form, as explained earlier. Therefore, their written Tamil vocabulary could be smaller than their written English vocabulary. This could also explain to some extent why the mean score of the Tamil X-lex was much lower than the mean score of the English X-lex; while Tamil is their dominant spoken language, English is the medium of instruction, most written text is in English, and English does not present the additional challenge of diglossia.

Additionally, the method of administration of the test could have affected the test score. In other studies, with children, the test was orally administered (Daller and Ongun 2018) in order to prevent confusion with unfamiliar spellings. In the present study, the children were given copies of the test and asked to circle the words, which may have led to lower scores. For the Tamil test in particular, this could have resulted in the children marking fewer words than they actually knew due to the effect of diglossia.

The scores of the C-test did not correlate with the collocation reading times either. It is not clear why we did not observe this correlation, but it should be noted that a recent meta-analysis highlighted that more work is needed to understand the construct of the C-test and the extent to which it reflects different aspects of language processing (McKay et al. 2021). However, there was still a correlation between the X-lex English score and the C-test scores, showing a correspondence between vocabulary size and language proficiency.

5. Implications of the Findings

From a pedagogical perspective, the overarching finding that the L1 plays a significant role in the processing of L2 collocations for young learners suggests that the L1 should not be ignored when it comes to teaching of collocations. It is well established that knowledge of collocations is essential for progression in proficiency levels, and integrating them into vocabulary teaching and learning is an important step in this direction. Teachers, as well as developers of learning materials, should give special attention to L2-only collocations. Teachers, wherever possible, should be aware of the L1 and whether L2 collocations exist in the L1. This is not feasible in contexts where children may have multiple L1s, but in

contexts where all children have the same L1, this can be used to the benefit of both the children and the teacher.

Theoretically, the findings lend support to the developing understanding of dual activation in the bilingual mental lexicon. Despite structural differences, the latest versions of models such as the BIA model (Dijkstra and Van Heuven 1998) and the Multilink model (Dijkstra and Rekké 2010) support the notion that dual activation occurs when a bilingual is presented with a lexical item, although this varies with individual proficiency and the translatability of the item in question. With the contribution of results from the increasing number of experimental studies on the processing of formulaic language, more detailed frameworks and theories can be developed.

6. Limitations of the Study and Recommendations for Future Research

The study had a number of limitations, which must be acknowledged. First, word length and frequency, which are known to influence reading times, were not controlled in the design of the stimuli. To address this, we checked the effects of word length and frequency as follows:

Regarding word frequency, it is established in vocabulary acquisition research that higher-frequency words are processed faster than lower-frequency words due to their increased exposure. It was found that the words with shorter reading times in the study had higher frequencies, supporting previous findings. However, it is worth noting that a study by Schmitt and Underwood (2004) found that the frequency of formulaic language had a significant effect on the processing speed of first-language speakers, but this effect was not observed in bilingual speakers. This discrepancy could be due to the frequency index used, which might not be suitable for bilingual speakers.

Regarding word length, it has been found that shorter words tend to have shorter reading times compared to longer words, which was also true in this study. In future studies, it is important to control for length and frequency to ensure that this does not interfere with analyzing the congruency effect.

Second, the use of self-paced reading as the methodology may have masked subtle differences in reading behavior and, in addition, may compromise natural reading behavior (Paape and Vasishth 2021). Eye-tracking could provide more insights into more natural reading (including regressions to reread parts of text) and should be considered in future research. Third, more objective measures of vocabulary knowledge (rather than relying on children's own judgments) might have been preferable, especially given the issue with diglossia in Tamil highlighted above. Similarly, while the C-test is widely used in the L2 literature, separate measures of vocabulary, grammar, and other components of language proficiency may have given a more detailed picture of children's proficiency (in both languages).

Author Contributions: Conceptualization, R.L. and H.J.; Methodology, R.L.; Software, R.L.; Validation, R.L.; Formal analysis, R.L.; Investigation, R.L.; Resources, R.L.; Writing—original draft, R.L.; Writing—review and editing, H.J.; Supervision, H.J. and M.D.; Project administration, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Felix Scholarship awarded by the University of Reading.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the University of Reading Ethics Committee (Date of approval: 2 March 2017).

Informed Consent Statement: Informed consent was obtained from all parents of all subjects involved in the study since the subjects themselves are below the age of consent.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Armon-Lotem, Sharon, Susan Joffe, Hadar Abutbul-Oz, Carmit Altman, and Joel Walters. 2014. Language exposure, ethnolinguistic identity and attitudes in the acquisition of Hebrew as a second language among bilingual preschool children from Russian-and English-speaking backgrounds. *Input and Experience in Bilingual Development* 13: 77–98.

Ayyar, Rajamalai Vaidyanatha. 1993. Educational Planning and Administration in India: Retrospect and Prospect. *Journal of Educational Planning and Administration* 7: 197–214.

Babaii, Esmat, and Hasan Ansary. 2001. The C-test: A valid operationalization of reduced redundancy principle? *System* 29: 209–19. [CrossRef]

Babatsouli, Elena, and Elena Nicoladis. 2019. The acquisition of English possessives by a bilingual child: Do input and usage frequency matter? *Journal of Child Language* 46: 170–83. [CrossRef] [PubMed]

Bonk, William, and Alice F. Healy. 2005. Priming effects without semantic or associative links through collocation. Paper presented at the 46th Annual Meeting of the Psychonomic Society, Toronto, ON, Canada, November 10–13; pp. 10–13.

Bultena, Sybrine, Ton Dijkstra, and Janet G. Van Hell. 2015. Language switch costs in sentence comprehension depend on language dominance: Evidence from self-paced reading. *Bilingualism: Language and Cognition* 18: 453–69. [CrossRef]

Carrol, Gareth, and Kathy Conklin. 2020. Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech* 63: 95–122. [CrossRef]

Carrol, Gareth, Kathy Conklin, and Henrik Gyllstad. 2016. Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition* 38: 403–43. [CrossRef]

Conklin, Kathy, and Ana Pellicer-Sánchez. 2016. Using eye-tracking in applied linguistics and second language research. *Second Language Research* 32: 453–67. [CrossRef]

Daller, Michael, and Zehra Ongun. 2018. The threshold hypothesis revisited: Bilingual lexical knowledge and non-verbal IQ development. *International Journal of Bilingualism* 22: 675–94. [CrossRef]

Dearden, John. 2014. *English as a Medium of Instruction—A Growing Global Phenomenon*. London: British Council.

Dijkstra, Ton, and Steven Rekké. 2010. Towards a localist-connectionist model of word translation. *The Mental Lexicon* 5: 401–20. [CrossRef]

Dijkstra, Ton, and Walter J. Van Heuven. 1998. The BIA model and bilingual word recognition. In *Localist Connectionist Approaches to Human Cognition*. London: Psychology Press, pp. 189–225.

Ding, Chen, and Barry L. Reynolds. 2019. The effects of L1 congruency, L2 proficiency, and the collocate-node relationship on the processing of L2 English collocations by L1-Chinese EFL learners. *Review of Cognitive Linguistics* 17: 331–57. [CrossRef]

Dörnyei, Zoltán, and Lucy Katona. 1992. Validation of the C-test amongst Hungarian EFL learners. *Language Testing* 9: 187–206. [CrossRef]

Durrant, Philip, and Norbert Schmitt. 2010. Adult learners' retention of collocations from exposure. *Second Language Research* 26: 163–88. [CrossRef]

Eckes, Thomas, and Rüdiger Grotjahn. 2006. A closer look at the construct validity of C-tests. *Language Testing* 23: 290–325. [CrossRef]

Fang, Nan, and Ping Zhang. 2021. L1 congruency, word frequency, collocational frequency, L2 proficiency, and their combined effects on Chinese–English bilinguals' L2 collocational processing. *International Journal of Bilingualism* 25: 1429–45. [CrossRef]

Gargesh, Ravinder. 2006. Language issues in the context of higher education in India. In *Symposium on Language Issues in English-Medium Universities across Asia*. Pok Fu Lam: University of Hong Kong, pp. 8–9.

Grotjahn, Rüdiger. 2022. Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* 4: 211–25.

Harsch, Claudia, and Johannes Hartig. 2015. What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly* 12: 333–62. [CrossRef]

Henriksen, Birgit. 2013. Research on L2 learners' collocational competence and development–a progress report. In *L2 Vocabulary Acquisition, Knowledge and Use*. Edited by Camilla Bardel, Christina Lindqvist and Batia Laufer. Morrisville: Lulu.com, pp. 29–56.

Hiebert, Elfrieda H., Judith A. Scott, Ruben Castaneda, and Alexandra Spichtig. 2019. An analysis of the features of words that influence vocabulary difficulty. *Education Sciences* 9: 8. [CrossRef]

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Ernst Klett Sprachen. Cambridge: Cambridge University Press.

Jiang, Nan. 2000. Lexical representation and development in a second language. *Applied Linguistics* 21: 47–77. [CrossRef]

Kim, Soo Hyon, and Ji Hyon Kim. 2012. Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *Tesol Quarterly* 46: 831–41. [CrossRef]

Kroll, Judith F., and Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33: 149–74. [CrossRef]

Lindsay, Shane, and Gareth M. Gaskell. 2010. A complementary systems account of word learning in L1 and L2. *Language Learning* 60: 45–63. [CrossRef]

McKay, Todd H., Yasser Teimouri, Aysenur Sağdıç, Bradford Salen, Derek Reagan, and Margaret E. Malone. 2021. The cagey C-test construct: Some evidence from a meta-analysis of correlation coefficients. *System* 99: 102526. [CrossRef]

McLaughlin, Barry. 1995. *Fostering Second Language Development in Young Children: Principles and Practices*. Berkeley: UC Berkeley: Center for Research on Education, Diversity and Excellence. Available online: https://escholarship.org/uc/item/23s607sr (accessed on 23 September 2024).

McLaughlin, Barry, ed. 2013. *Second Language Acquisition in Childhood: Volume 2: School-Age Children*. London: Psychology Press.
Meara, Paul M., and James Milton. 2003. *X-lex: The Swansea Levels Test*. Berkshire: Express Publishing.
Milton, James. 2006. X-Lex: The Swansea vocabulary levels test. In *Proceedings of the 7th and 8th Current Trends in English Language testing (CTELT) Conference*. Dubai: TESOL Arabia, vol. 4, pp. 29–39.
Milton, James. 2007. Lexical profiles, learning styles and the construct validity of lexical size tests. In *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, pp. 47–58. [CrossRef]
Murphy, Victoria A. 2014. *Second Language Learning in the Early School Years: Trends and Contexts-Oxford Applied Linguistics*. Oxford: Oxford University Press.
Northbrook, Julian, and Kathy Conklin. 2019. Is what you put in what you get out?—Textbook-derived lexical bundle processing in beginner English learners. *Applied Linguistics* 40: 816–33. [CrossRef]
Paape, Dario, and Shravan Vasishth. 2021. When Nothing Goes Right, Go Left: A Large-Scale Evaluation of Bidirectional Self-Paced Reading. Available online: https://osf.io/d7pvz (accessed on 20 September 2024).
Paquot, Magali, and Hubert Naets. 2015. Adopting a relational model of co-occurrences to trace phraseological development. Paper presented at the Learner Corpus Research, Nijmegen, The Netherlands, September 11.
Paradis, Johanne, Elena Nicoladis, Martha Crago, and Fred Genesee. 2011. Bilingual children's acquisition of the past tense: A usage-based approach. *Journal of Child Language* 38: 554–78. [CrossRef]
Peirce, Jonathan W. 2007. PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods* 162: 8–13. [CrossRef]
Pliatsikas, Christos, and Theodoros Marinis. 2013. Processing of regular and irregular past tense morphology in highly proficient second language learners of English: A self-paced reading study. *Applied Psycholinguistics* 34: 943–70. [CrossRef]
Ponnuchamy, Gitarani. 2012. School English-as-a-Second-Language Experiences of Students at Tertiary Institutions in Tamil Nadu, India: A Phenomenological Study. Ph.D. thesis, University of Phoenix, Phoenix, AZ, USA.
Pritchett, Lena K., Jyotsna Vaid, and Sumeyra Tosun. 2016. Of black sheep and white crows: Extending the bilingual dual coding theory to memory for idioms. *Cogent Psychology* 3: 1135512. [CrossRef]
Schmitt, Norbert. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Berlin and Heidelberg: Springer.
Schmitt, Norbert, and Geoffrey Underwood. 2004. Exploring the processing of formulaic sequences through a self-paced reading task. In *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins Publishing Company, pp. 173–89.
Siyanova, Anna, and Norbert Schmitt. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review* 64: 429–58. [CrossRef]
Siyanova-Chanturia, Anna, and Diana V. L. Sidtis. 2018. What online processing tells us about formulaic language. In *Understanding Formulaic Language*. London: Routledge, pp. 38–61.
Wolter, Brent, and Henrik Gyllstad. 2011. Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics* 32: 430–49. [CrossRef]
Wolter, Brent, and Henrik Gyllstad. 2013. Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition* 35: 451–82. [CrossRef]
Wolter, Brent, and Junko Yamashita. 2018. Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition* 40: 395–416. [CrossRef]
Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Port Chester: Cambridge University Press, p. 10573.
Yamashita, Junko, and Nan Jiang. 2010. L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *Tesol Quarterly* 44: 647–68. [CrossRef]
Yip, Virginia, and Stephen Matthews. 2000. Syntactic transfer in a Cantonese–English bilingual child. *Bilingualism: Language and Cognition* 3: 193–208. [CrossRef]
Zulfikar, Zulfikar. 2019. Rethinking the use of L1 in L2 classroom. *Englisia: Journal of Language, Education, and Humanities* 6: 42–51. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Improving Classroom Teaching and Learning of Multi-Word Expressions for Conversational Use Through Action Research with Learner Feedback

Haidee Thomson

Junior College Department, Hokusei Gakuen University, Sapporo 004-8631, Hokkaido, Japan; h-thomson@hokusei.ac.jp

Abstract: Multi-word expressions make up a large proportion of the English language and particularly spoken language. Using multi-word expressions can assist with the impression of fluency, making them useful for language learners to know and use. However, proven methods for teaching this language phenomenon are required, so that learners can easily use multi-word expressions in their conversations. The purpose of the study was to examine the efficacy of a fluency workshop focused on multi-word expression use in conversation and to determine the most appropriate implementation for the Japanese context. An action research structure was used over three iterations of the fluency workshop, learner feedback and teacher observations were used to make improvements. Learner feedback regarding the usefulness of each activity for learning English was compared between the original cohort and subsequent cohorts. The results showed significant differences in levels of perceived usefulness for activities where improvements were made, but also for some activities where no specific improvement was made, suggesting that teaching improves through practice. Pedagogical implications include maximising the time on task via clear instructions, providing visual time constraints, and offering scaffolding to support the use of multi-word expressions when recall seems beyond a learner.

Keywords: multi-word expressions; conversational fluency; action research; role play; dictogloss; time on task

1. Introduction

The ubiquitous presence of multi-word expressions in English, especially in spoken language, suggests that language learners need to learn and know these common sequences (Vilkaitė 2016). The term, "multi-word expression", refers to word sequences that are often formulaic including "…(semi-) fixed recurrent phrases, … multi-word verbs (put up with), …speech formulae (What's up), lexical bundles (in the middle of), and other types" (Siyanova-Chanturia and Martinez 2015, p. 549). Multi-word expression use has been increasingly linked to speaking fluency in recent years (Boers et al. 2006; Tavakoli and Uchihara 2020), suggesting that the use of frequent multi-word expressions offers a boost to fluency for second-language and foreign-language learners (Wulff 2019). In search of an effective classroom approach to teach multi-word expressions, this paper will review the literature regarding how multi-word expressions can be learned. A classroom-based approach to learning multi-word expressions (e.g., I think I will, would you like to), is then reviewed and set up as an action research investigation, informed by learner feedback and teacher observations to suggest improvements for the learning of multi-word expressions in the EFL classroom in Japan.

Frequency of encounter is key to learning new vocabulary items (Webb 2007), and this is similar for multi-word expressions (Peters et al. 2023; Webb et al. 2013). Many English-language learners in Japan have limited exposure to the language in their daily life, and although English is taught in schools, the proficiency in Japan is low, ranking

87 out of 113 countries (EF English Proficiency Index 2023). The opportunities to speak are limited, even within the English classroom in Japan (Koizumi et al. 2022). Therefore, without intentional and repeated exposure to multi-word expressions, learners are unlikely to have enough encounters with the new language items to remember them for future use. Furthermore, the individual words that make up multi-word expressions are often already known by learners thanks to their high frequency, with many comprising multiple-function words, for example "is on the right". As a consequence, learners may fail to notice a multi-word expression as a new item to learn due to the non-salient nature of the component words (Jiang 2009). The noticing hypothesis (Schmidt 2001; Schmidt 1990) proposes that language items must be consciously noticed in order to be learnt, while the weaker version states that more noticing leads to more learning. There is also evidence to suggest that explicit noticing may not be required in the acquisition of transparent multi-word expressions (see Szcześniak 2024).

Textbooks often present exercises to raise learners' awareness of multi-word expressions without prior modeling of correct usage. Typical exercises include gap-fill, multi-choice exercises, and matching the parts of the expression together. Strong and Boers (2019) found that learning phrasal verbs is more successful when correct usage examples are given, and the exercises are used for retrieval. In 25% of cases, where correct usage was not modelled, learners were repeating their errors one week later despite receiving corrective feedback. Toomer and Elgort (2019) found that for advanced learners, bolding collocations within a 500-word text where the collocation appears three times can enhance explicit knowledge, while reading the same text with no bolding can improve implicit knowledge of the collocations.

Previous research suggests that the active noticing of multi-word expressions in class leads to greater recycling of multi-word expressions from a test prompt (Boers et al. 2006). However, in a similar study (Stengers et al. 2010), when the test prompt was offered only in the L1, the participants did not use more multi-word expressions than the control group, suggestive that the noticing strategies used in class had not resulted in the ability to use any more multi-word expressions. In light of this, Boers and Lindstromberg (2012) argue that mere noticing or awareness raising is not enough if encounters are few and far between, leaving little or no memory trace. A more recent study (Boers et al. 2023) again drew learners attention to the phenomenon of multi-word expressions within classroom texts with additional guidance on how to use dictionaries and corpora to determine whether a string of words were indeed a multi-word expression, after 11-weeks the experimental group were writing essays using more accurate and wide-ranging multi-word expressions than the control group.

Lin (2019) argues that to learn multi-word expressions through reading alone would be to miss important features of their spoken prosody. Most of the multi-word expressions that L1 speakers know are likely learnt through hearing them in conversations. Young L1 learners will pick up the melody of expressions before the specific words and often use them with the correct prosody, even when they have yet to form the words (Peters 1983). Multi-word expressions can have specific prosodic rhythms which set them apart, so it seems important for L2 learners to hear multi-word expressions and learn their prosody when they learn them (Lin 2019). One way to learn, which involves hearing the prosody, is through watching online videos. Majuddin et al. (2021) found that learners in Malaysia could recall multi-word expressions better when they watched a 20 min video with captions in English twice, suggestive that incidental learning can occur if there is repetition.

Repetition is golden when it comes to language learning. However, repetition without variation becomes difficult to sustain, especially in a classroom. Some successful examples include Wray (2004) where one motivated TV presenter memorised the script to present a cooking programme on TV in the Welsh language (which was relatively new to her). Another relatively successful example was Wray and Fitzpatrick (2010) where learners chose utterances that they anticipated would be useful and had to memorise them. They then tried having conversations where they relied on these utterances and, depending on the learner, found

them confidence instilling and helpful for sustaining their conversations while also feeling the constraints of not knowing more. It therefore seems reasonable to expect that learning will be expediated by deliberate noticing accompanied by practice using the multi-word expression (to create and consolidate strong memory traces), particularly when learners have limited exposure to the foreign language, as is the case for learners of English in rural Japan.

In light of the potential for improved speaking fluency, classroom models for teaching multi-word expressions to low proficiency EFL learners are needed. Wood (2009), reported a promising model, when he reported a "fluency workshop" which gave learners multiple and varied activities to notice and repeatedly use multi-word expressions in the classroom. The intervention was reported to increase fluency, however only at a case-study level in Canada with a Japanese learner of English. The fluency workshop activities were well-known activities staged from input to automatisation. The workshop included activities such as shadowing, dictogloss (Wajnryb 1990) (text reconstruction from audio), and mingle jigsaw (Wood 1998) (which involves memorising and sharing multi-word expressions). After these activities came practice and production, with activities such as 4/3/2 decreasing the time speaking on a topic with alternating partners (Arevart and Nation 1991; Boers and Thai 2017; Maurice 1983), and a free stage of speaking using the newly learnt expressions. The logical flow from input to practice building up to freer production with a variety of activities using the same multi-word expressions appealed to me as it is very reminiscent of Nation's (2007) four strands approach to fluency. The four strands start with meaning-focused input, moving to meaning-focused output, language-focused learning, and finally fluency activities. The progression provides learners with the opportunity to speak at a higher level of fluency due to the progression from preceding activities where they have built familiarity with the language. This progression, therefore, is likely to tap into the learners' proximal zone of development (Vygotsky 1978), a concept where tasks that are just outside of a learner's independent abilities are made possible through support. The interactive nature of the activities also align with the socio-cultural theory that we learn language better when we use it to interact with each other (Vygotsky 1978).

The fluency workshop "floods" the learner with multi-word expressions much like the graded readers in the research by Webb et al. (2013) and Peters et al. (2023) where graded readers were "flooded" with particular multi-word expressions, and more encounters were associated with greater incidental learning. What is different here, however, is that learners are given multiple activities flooded with prompts to use the multi-word expressions (find them, repeat them, say them, recall them, roleplay with them etc.) rather than simply hear them while reading. From a teacher's point of view the fluency workshop seemed practical and versatile. It also appears able to be adapted to any conversation context that is desired because once there is a conversation with the multi-word expressions, the activities can easily be created to go with it. Therefore, I adapted the fluency workshop to use in my EFL classes in Japan, with a focus on conversational interaction in travel-themed scenarios. The travel theme was selected since previous cohorts of students had responded to surveys about what they thought would be useful to learn in class, and using English for travel was a prominent response. The adaption into a dialogue version for English learners in Japan was new. I therefore sought to collect evidence through action research regarding the efficacy and appropriateness of the intervention in this form and whether it could be improved for the specific learning context in Japan.

In action research, the researcher is often also the teacher, who identifies a classroom problem and then experiments with interventions to solve the problem (Burns 2010). In this case, I was the researcher and teacher, and I wanted to develop an intervention to address the lack of speaking fluency that my university-level learners of English displayed. The standard teaching practice was to use a textbook and work through each unit with limited opportunity for repeated practice with multi-word expressions. However, due to the links made in the research between the use of multi-word expressions and fluency, I felt that helping learners to use common multi-word expressions might help boost their speaking fluency. Action research follows an iterative cycle of question, intervention, and reflection

leading to the next cycle. Following this cycle is known to lead to improved teaching practice and is used by many research-orientated teaching practitioners to improve their teaching. The research questions were therefore as follows:

Research question 1: To what extent is a dialogue version of the fluency workshop effective for learning multi-word expressions?
Research question 2: What aspects of the classroom experience of the fluency workshop can be improved/made more appropriate for the Japanese context?

2. Materials and Methods

The specific action research cycle deployed here was to determine whether the fluency workshop intervention was effective and appropriate, and whether it could be improved for the Japanese context. The fluency workshop was run with three different cohorts of students over three semesters. I collected feedback from the participants regarding their experience of the activities via surveys and focus group interviews. I also took note of my teacher/researcher observations and reflections about whether improvements could be made to the classroom intervention. This feedback was then used to make improvements to the fluency workshop activities, and then the fluency workshop would be run with the second cohort of students. Feedback would then again be collected and reflected on, resulting in further improvements if required before running the fluency workshop again with the third cohort of students. The multi-word expressions, intervention activities, tests, and feedback surveys used will be described next.

In order to determine whether the fluency workshop intervention was effective, learner knowledge, use, and fluency with 30 multi-word expressions were measured using tests before and after the intervention. To measure knowledge, a cued recall test with the first letters of the words for the multi-word expressions included in a context sentence was administered via Moodle on the computers in the classroom. Then, in partners, the participants roleplayed a conversation which was recorded on the computers using Audacity in order to measure their use of the multi-word expressions in conversations along with fluency. The interactions during the activities in the fluency workshop were observed by the teacher/researcher and also reflected on by the student participants via surveys and focus group interviews after completing the fluency workshop. Feedback from the participants and teacher observations were used to make improvements to the classroom activities during a second and third iteration of the intervention with fresh participants/students experiencing the workshop each time.

Ethics approvals were applied for and granted at the research and teaching institutions. The students were informed about the nature of the research and asked to complete consent forms at the beginning of each study. The data reported are only from those who gave their informed consent. The participants were recruited from intact EFL classes (convenience samples) in an engineering university in Japan. The participants were only required to take two English courses as part of their degree. Self-reported TOEIC scores ranged from 255 to 600, equivalent to a false beginner to low intermediate A2-B1 CEFR range, there was no significant difference in this proficiency measure between the three cohorts of experimental groups according to the non-parametric Kruskal–Wallis test: $H(2) = 1.573$, $p = 0.455$.

Each fluency workshop was conducted over six 90 min class sessions using three units, intended to raise awareness, knowledge, use, and fluency with 30 multi-word expressions of four words. The materials for the experimental condition had travel-related themes that the students might experience and therefore consider useful (ordering food at a cafe, asking for directions on the street, and booking a hotel room). Each unit was built around a model dialogue developed by the researcher containing ten common multi-word expressions (transparent in nature). A list of the specific MWEs used along with frequency information can be seen in Appendix A. The multi-word expressions were checked against the spoken sub-corpus of COCA (Davies 2013) to confirm their common usage. The learners were familiarised with one round of the fluency workshop activities in class using different multi-word expressions before the study began.

The experimental group participated in the following fluency workshop activities based on the Wood (2009) model but adapted to a dialogue model. The activities started with input and progressively built knowledge and practice with support, leading to fluency activities where the multi-word expressions were used in roleplay conversations. The concept behind this progression is that after noticing the multi-word expressions in the context of the model dialogue, practising the pronunciation and rhythm, trying to recall the multi-word expressions from memory and using them in roleplay, that the learners would become familiar with the multi-word expressions and then try to use them in their own conversations. The activities were as follows:

(1) Listening to the dialogue as a class and answering gist questions in pairs;
(2) Marking pauses—the students listened to the audio as a class and marked pauses on the transcript (this activity was removed for the second and third cohorts);
(3) Phrase instruction—the teacher showed the multi-word expression on the class screen and explained the meaning and function, and the students then found and marked the multi-word expression in their transcript and were encouraged to take notes;
(4) Shadowing—the students listened to the audio on computer headsets repeating the words just after the speaker four times to learn the pronunciation and prosody/rhythm;
(5) Dictogloss (Wajnryb 1990) (text reconstruction from audio)—the students covered their transcripts and listened to the audio once as a class without taking notes, then listened a second time while taking notes and sharing with a partner, after which they listened again and shared, and listened a fourth time before sharing notes with a small group and attempting to reconstruct the dialogue together;
(6) Mingle jigsaw (Wood 1998)—the students were each given a multi-word expression on a piece of paper to memorise; they had to write it down and cover it so others could not see it. Papers were collected and then the students exchanged their multi-word expressions orally with others; they walked around sharing their multi-word expressions and returned to their desks to write down each multi-word expression. They were not permitted to write while listening to a speaker, and the goal was to find all 10 multi-word expressions;
(7) Roleplay with notes—the students lined up facing a partner and roleplayed the dialogue while referring to notes; they rotated with partners three times and then switched roles and rotated with another three partners;
(8) Decreasing-time roleplay without notes (based on 4/3/2 (Arevart and Nation 1991; Boers and Thai 2017; Maurice 1983)—the students roleplayed with a time limit that decreased with each partner change. They roleplayed with three different partners and then switched roles and rotated with another three partners;
(9) Role-play recording—the students sat with a partner in front of a computer and audio recorded their roleplay without looking at notes. The students listened back to their recording and reflected with a worksheet on their speaking fluency (added from the second cohort onwards);
(10) A related situation roleplay—the students chose a situation from several options that were similar to the model and then roleplayed ad lib with the encouragement to use the multi-word expressions if they could.

Participant feedback from the surveys and focus groups along with the teacher observations were considered and adjustments/improvements were made to the activities for the second and third cohorts as shown in Table 1.

One unit took about 160 min to complete. All three units along with pre- and post-testing were completed within six class sessions. During the final unit, the students were surveyed about their experience with each activity after they completed it. A control group which received no manipulation was included with the first and third cohorts in order to provide a baseline to consider whether any improvement within the experimental groups was due to the treatment or some other reason (see Loewen and Plonsky (2016, p. 34) for the definition of a control group). The control group was not deliberately exposed to the multi-word expressions (other than during the pre-test). They learned English for

the same number and length of class sessions; however, they continued with their usual focus on engineering topics through linked skill activities intended to improve their fluency progressively through listening, reading, writing, and finally speaking on each topic.

Table 1. Adjustments made to the fluency workshop activities after the first cohort.

Activity	Improvement
Shadowing	Added an explicit instruction to pause/replay as desired
Dictogloss	Changed to a small-group activity from the start with a visible countdown timer and a simplified worksheet
Mingle jigsaw	Added a visible countdown timer
Decreasing-time roleplay with no notes	Added a display of multi-word expressions, adjusted the timings, and added scaffolding for the hotel unit
Recording the roleplay	Added time and structure to listen to and reflect on the recording
Related situation roleplay	Reduced the scenario choices, displayed images, and added planning time in L1

In order to collect feedback from the participants about their experience with the classroom activities, a Likert scale survey was used. The survey asked the participants about how useful they perceived the activity to be for learning. The Likert scale for the first cohort had five response options: strongly disagree, disagree, neither disagree nor agree, agree, and strongly agree. For the second and third cohorts, the Likert scale was extended from five to six response options, replacing the neutral [neither agree nor disagree] with a choice of [slightly disagree] or [slightly agree]. In the first cohort, the tendency to choose the neutral response was strong, so extending to a six-point scale was intended to push the participants to show a preference. In order to compare their responses with those of the first cohort, the responses were analysed as a binary of either agree or disagree, and neutral responses were not included. The participants could also write comments about their experience with the activity. Furthermore, focus group interviews were conducted with several self-selected participants; these were audio recorded, and informative comments were noted to help further understand the learner experience of the activities. Additionally, the researcher/teacher made observations about how effective the learning activities seemed to be during the class sessions and how they could be improved going forward. The notes from both the focus groups and the teacher observations were used to complement the activity feedback survey results with more in-depth qualitative explanation.

The second cohort had fewer participants and lacked a control group; it served as the testing ground for the changes to the intervention based on the feedback from the first cohort using the action research feedback cycle. The third cohort of participants essentially experienced the same as the second cohort with minor improvements. Therefore, the survey responses regarding the intervention from the second and third cohorts were grouped together for the analysis resulting in two groups: first cohort, and second and third cohorts. I hypothesised that the responses from the participants from the latter cohorts would be more favourable for the activities where the changes had been made as part of the action research because in theory, the changes would improve the learning experience.

3. Results

In order to answer the first research question regarding whether a dialogue version of the fluency workshop is effective for learning multi-word expressions, the learning results from the intervention, such as multi-word expressions knowledge, use, and fluency were measured with pre- and post-intervention testing. When the results are compared between the first and third cohorts, it is clear that the results were better for the third cohort, which is likely due to the improvements made to the implementation as part of the action research cycle. The results from the first and third cohorts have previously been published (see Thomson 2018; Thomson et al. 2023); therefore comparisons will be made

between these two cohorts using Cohen's d effect size measure. For the first cohort, the quantitative results showed "...productive knowledge of target expressions had increased significantly more for the experimental group (n = 44) M = 7.63, SD = 4.12, than for the control group (n = 29) M = 2.55, SD = 3.01, $t(71) = -5.71$, $p = 0.000$, d = 1.41" (Thomson 2018, p. 39). However, there was no significant increase in the use of the multi-word expressions in conversations or fluency when compared with the control group. The results for the third cohort were reported in Thomson et al. (2023) and showed a stronger increase in multi-word expression knowledge for the experimental group (n = 65) M = 8.71, SD = 3.64, than for the control group (n = 51) M = 2.50, SD = 2.92, $t(114) = 9.93$, $p < 0.001$, d = 1.88. The third cohort also showed a significant increase in the use of the multi-word expressions when compared to the control group, which was not present for the first cohort (Thomson et al. 2023). These results suggest that the changes made to the activities in response to the student feedback and teacher observations improved the learning outcomes for the multi-word expressions between the first and third cohorts. These comparisons with the control group suggest that the dialogue version of the fluency workshop is effective for learning multi-word expressions.

To answer the second research question about what aspects of the fluency workshop can be improved/made more appropriate for the Japanese context, the learners' feedback via the activity feedback surveys and focus group interviews, along with the teacher/researcher observations are considered. The learners' response to the introduction activity of listening to the audio and answering the gist questions were not surveyed as it was a fairly standard classroom activity setting the scene for the following activities. Additionally, marking pauses was removed after the first cohort because the dialogue short turn format meant that pauses were few and hard to notice. The focus group participants from the first cohort reported marking pauses where they thought that they would go since they could not distinguish many. Pearson chi-square and Fishers exact test (when numbers were below the expected limits for chi-square) were used to compare the learner activity feedback responses between the first cohort and latter cohorts (second and third cohorts). The participants responded in the fifth and sixth class sessions, after completing three units, so they had experienced each activity type three times before completing the survey (the numbers vary due to the variations in attendance and completion of the surveys). The comparisons between the responses from the first cohort and the latter cohorts regarding the remaining activities are reported below. Reference is made to the participant comments where relevant from the focus group interviews and comment sections of the survey.

The phrase instruction was not altered between the first and latter cohorts as 91.2% of the participants from the first cohort responded that it was a useful activity. Table 2 shows the response numbers and percentages for the phrase instruction. As expected, Fisher's exact test showed no significant difference between the first and latter cohort responses with both groups generally agreeing that the phrase instruction was useful for learning English (two-tailed $p = 0.755$).

Table 2. Comparison of the participant responses for the phrase instruction.

Phrase Instruction	Not Useful	Useful
First cohort	3 (8.8%)	31 (91.2%)
Second and third cohorts	11 (12.6%)	76 (87.4%)

For the shadowing activity, some focus group interview participants shared that they had not realised that they could stop and replay the audio as they liked; therefore, an explicit instruction was added for the latter cohorts to stop, start, and replay the audio as they needed. Initially, I had doubted the usefulness of shadowing, but the feedback from the focus group interview participants convinced me that the activity was appreciated by the students for learning the pronunciation and rhythm of the dialogue. It is therefore possible that I gave more confident and knowledgeable instructions in the latter iterations

of the workshop, which may have also improved the activity. Table 3 shows that 85.7% of the latter cohort agreed that shadowing was useful for learning English, and this was significantly more than the original cohort according to a Pearson chi-square analysis X2 (1, N = 122) = 4.94, $p = 0.026$.

Table 3. Comparison of the participant responses for shadowing.

Shadowing	Not Useful	Useful
First cohort	14 (31.1%)	31 (68.9%)
Second and third cohorts	11 (14.3%)	66 (85.7%)

Dictogloss was changed in response to feedback and observations from the first cohort. In particular, I observed that the small groups were not sharing well, and I attributed this to the initial work performed in pairs, which seemed to lead to uncomfortableness when expanding into a group of four for the final reconstruction. There also appeared to be hesitation to be the first to start sharing. Therefore, for the latter cohorts, the learners worked in small groups of four from the beginning of the activity to compare notes and reconstruct the text, with a five-minute timer for the final reconstruction. The students appeared more willing to share openly when having this arrangement from the beginning and worked to complete the task faster under the time pressure. The dictogloss worksheet was also simplified with numbers added to the instructions and also to the dialogue lines for reconstruction, providing the learners with a short reference system for comparing and sharing in their groups. Table 4 shows that 76.4% of the latter group agreed that the dictogloss activity was useful for learning English, and this was significantly more than of the original group: X2 (1, N = 117) = 5.39, $p = 0.020$.

Table 4. Comparison of the participant responses for dictogloss.

Dictogloss	Not Useful	Useful
First cohort	13 (46.4%)	15 (53.6%)
Second and third cohorts	21 (23.6%)	68 (76.4%)

Regarding the mingle jigsaw activity, the participants in the focus group interviews from the first cohort commented that it was hard to approach others in the class whom they were unfamiliar with, and some people did not seem interested in sharing. In the latter cohorts, a five-minute countdown timer was therefore displayed to motivate the action to collect all ten expressions within the time limit, much like a race. Table 5 shows that 79.1% of the participants from the latter studies agreed that the mingle jigsaw was useful for learning English. The Pearsons chi-square test showed that this was a statistical increase from the first cohort X2 (1, N = 119) = 8.87, $p = 0.003$.

Table 5. Comparison of the participant responses for the mingle jigsaw.

Mingle Jigsaw	Not Useful	Useful
First cohort	16 (48.5%)	17 (51.5%)
Second and third cohorts	18 (20.9%)	68 (79.1%)

There were no changes made to the roleplay activity because 72.5% of the first cohort responded that the roleplay was useful for learning English; furthermore, the focus group interview participants were satisfied with the roleplay as it was. Table 6 shows that 89.1% of the responses for the latter cohorts perceived the roleplay to be useful for learning English, which is statistically more than of the first cohort: X2 (1, N = 132) = 5.74, $p = 0.016$. Despite making no changes to the activity, it is entirely possible that, as the teacher, my instructions and confidence had improved with the experience and practice from the first cohort which therefore improved the learners' perception about how useful the roleplay was.

Table 6. Comparison of the participant responses for the roleplay.

Roleplay	Not Useful	Useful
First cohort	11 (27.5%)	29 (72.5%)
Second and third cohorts	10 (10.9%)	82 (89.1%)

The decreasing-time roleplay was intended to encourage the learners to feel time pressure and speak faster. In the first cohort, the time increments for the decreasing-time roleplay started at 120 s, decreasing to 90 s, and finally 60 s. The feedback from the focus group participants revealed that 120 s and also 90 s were longer than necessary. Therefore, the time increments were shortened to 100, 80, and 60 s for the café and directions roleplay. However, the focus group participants pointed out that the model dialogue for the hotel unit was longer and more challenging, and they struggled to finish this one within the time constraints. Therefore, I shortened the model dialogue for the hotel unit and added a realia booking form to remind the learners about the details that they had to request from their conversation partner. The time increments for the hotel unit were set longer at 120, 100, and 80 s for the third cohort. As the teacher, I also observed the learners substituting words and phrases when they could not recall the multi-word expressions, since they had no notes to refer to. In order not to lose the opportunity to use the multi-word expressions during this activity, I decided to display the multi-word expressions on the class screen for reference for the second and third cohorts. Table 7 shows that 86% of the participants in the latter cohorts (significantly more than of the first cohort) responded that the decreasing-time roleplay was useful for learning English; X^2 (1, N = 121) = 10.00, $p = 0.002$.

Table 7. Comparison of the participant responses for the decreasing-time roleplay.

Decreasing-Time Roleplay	Not Useful	Useful
First cohort	14 (40.0%)	21 (60.0%)
Second and third cohorts	12 (14.0%)	74 (86.0%)

The purpose of recording audio is usually so that you can listen back to it. In the first cohort, the learners recorded their roleplays but were not given time in class to listen back to the complete recording. As the teacher, I assumed (incorrectly) that the students would listen back to the recordings after class. Without listening back to your recording, this activity lacks usefulness for learning. The feedback from the focus group interviews revealed that few students listened back to their recordings despite having access. Therefore, I created a worksheet for the participants in the latter cohorts to report their reflections on their fluency while listening back to their recordings during class time. Table 8 shows that 86.4% of the respondents from the latter cohorts found the roleplay recording activity useful for learning English, statistically, more than of the first cohort; X2 (1, N = 125) = 13.14, $p < 0.001$.

Table 8. Comparison of the participant responses for the roleplay recording activity.

Roleplay Recording	Not Useful	Useful
First cohort	16 (43.2%)	21 (56.8%)
Second and third cohorts	12 (13.6%)	76 (86.4%)

In the first cohort, for the related situation roleplay activity, I observed the participants taking a disproportionate amount of time deliberating with their partner about which scenario to roleplay, leaving almost no time to roleplay. Therefore, I reduced the scenario options from four to two, and added photos on the class screen as scaffolding to help the learners imagine the scenario and start the roleplay faster. Also, I added one minute

of planning time so that the learners could discuss their intentions using their L1 before starting the roleplay. Table 9 shows that 86% of the respondents from the latter cohorts considered the related situation roleplay activity to be useful for learning English. Again, statistically, this was significantly more than the first cohort; X2 (1, N = 124) = 10.08, $p = 0.002$. In the next section, these results will be discussed in relation to previous research and teaching implications.

Table 9. Comparison of the participant responses for the related situation roleplay activity.

Related Situation Roleplay	Not Useful	Useful
First cohort	15 (39.5%)	23 (60.5%)
Second and third cohorts	12 (14.0%)	74 (86.0%)

4. Discussion

The first research question asked to what extent a dialogue version of the fluency workshop (inspired by Wood 2009) would be effective for learning multi-word expressions. The adaption and modification of the fluency workshop to a dialogue model for a Japanese EFL learning context, based on the results from this series of studies appear to increase learners' knowledge of multi-word expressions and use of multi-word expressions (Thomson 2018; Thomson et al. 2023). Another fluency-workshop-inspired study (McGuire and Larson-Hall 2018) also using a dialogue format reported fluency increases. However, in the current research, fluency did not appear to increase when compared with a control group. The difference between an increase or not in fluency could be related to the pre- and post-intervention speaking prompt; in the current set of studies, the learners were roleplaying a scenario, whereas in the McGuire and Larson-Hall (2018) study, the participants spoke about their personal likes and dislikes for the fluency measurement, which may have been easier to speak about versus an imagined scenario. Additionally, their participants had higher proficiency and were studying in the US, so they would likely have had more exposure to English use outside of the classroom than the EFL learners based in Japan in these studies. Perhaps using a monologue speaking task to measure fluency might give more insightful results in regard to fluency. Additionally, it could be argued that fluency in a conversation and fluency in a monologue are two different types of fluency as the conversational version requires some social skills to understand your interlocuter and respond appropriately while also encouraging continued dialogue. In any case, it would be interesting to test speech fluency via a single-person speaking task to ensure that increases in fluency are not missed due to the social dynamics involved in conversations.

The fluency workshop activities began with guided noticing of the multi-word expressions within the context of hearing and reading a model dialogue. The necessity of noticing transparent multi-word expressions has been questioned in the European context (Szcześniak 2024). While I would argue that these multi-word expressions are indeed transparent or fairly transparent, the activities in the fluency workshop preclude incidental learning and fairly explicitly draw attention to the multi-word expressions. Perhaps these particular multi-word expressions could be learnt simply through incidental encounters; however, the control group presents the result of leaving learning up to chance in the Japanese context, and the encounters with English are too few and far between to provide any guarantee of the learning of even fairly common multi-word expressions. Previous studies (Boers et al. 2006; Stengers et al. 2010) have found that noticing alone is unlikely to result in strong enough memory traces for learners to later recall the multi-word expressions without support. The fluency workshop intervention has gone beyond simply noticing and created stronger memory traces through practice as suggested by Boers and Lindstromberg (2012).

The fluency workshop activities included shadowing, which to some may seem like an unnecessary activity. However, the Japanese participants, with their L1, which sounds very different to English, insisted that shadowing was important because it helped them

to practise how to pronounce and say the multi-word expressions. The usefulness that the learners perceived for this activity supports the assertation that Lin (2019) makes for learning prosody when learning multi-word expressions.

The second research question asked what aspects of the classroom experience of the fluency workshop can be improved/made more appropriate for the Japanese context. The learners found the dictogloss (Wajnryb 1990) challenging but were helped by working to reconstruct the text in small groups of four people. The grouping was beneficial because if one person had not been able to take notes, someone else in the group would have taken some notes. Sharing in the group from the start worked better than changing from pairs to groups part way through. Also, the addition of a visible countdown timer for the final time to share and reconstruct the text appeared to help the learners overcome their fears about being the first to speak in their group. Similarly, the mingle jigsaw activity was difficult for the students to start as they had to decide whom to approach, and for some this resulted in holding back and not being as active as they had the potential to be. Simply displaying a visible five-minute countdown timer for the dictogloss and mingle jigsaw activities visibly sped up the initial interactions between the students as they were aware that they had limited time to complete a task. Raising the learners' awareness of the limited time available appeared to motivate the learners to interact with more vigour. It is possible that time pressure is more helpful with introverted learners, or those who are not so confident socially to approach others. I would also argue that the visual countdown keeps the task at the forefront, and conversations will be less likely to go off topic.

The time constraint also added a game-like element to the tasks, with some perhaps even seeing it as a race. This gamification of the activity appeared to speed up the students' activity, helping the students to increase their time on task (a concept which is generally associated with better learning outcomes) (Godwin et al. 2021). The visible time limit appeared to move the focus onto completing the task rather than feeling the fear of initiating a conversation. The use of a classroom timer to speed up the students' activity is not a new idea, and Wurtele and Drabman (1984) reported preschoolers speeding up their actions to tidy up their classroom after "beat the buzzer" was introduced. Using countdown timers has been credited with creating a sense of urgency in students to put in more effort and complete the task assigned (Olipas and Luciano 2020). Based on this experience, I would recommend using a countdown timer where teachers want to speed up student engagement and intensify engagement with the task.

Simplicity and the clarity of instructions and tasks was very important for quick transitions into each learning activity. For example, the instructions for the dictogloss were simplified from the original, and numbers were added on the worksheet to help the learners focus on what to do next. These number references likely also assisted with the communication between small-group members. During the reconstruction process, the learners could state a number reference to describe which part of the dialogue that they were talking about. Reducing the choice is another form of simplification to help the learners quickly move into the activity. For example, the related situation roleplay choices were reduced which led to faster decisions being made by students and likely resulted in more time interacting with the multi-word expressions. Simplifying the instructions or task helped the learners to spend more time using the multi-word expressions.

The provision of appropriate scaffolding helped the learners to activate their knowledge to use in interactions. For example, the related situation roleplay was not only simplified by reducing the scenario options, but photos aimed at activating the imagination, knowledge, and vocabulary for the scenario options were shown for the latter cohorts. The visual scaffolding appeared to help the learners imagine the scenario, remember the related words, and roleplay with more confidence. Both simplification and scaffolding helped to speed up the interaction and time on task, thereby increasing the time spent using the multi-word expressions for communication.

Further to the scaffolding point, an additional exposure to the multi-word expressions was added during the decreasing-time roleplay. Notes were hidden during this activity,

so if the learners were unable to recall the multi-word expressions, they would need to use other words to complete the roleplay. As the teacher, I observed the learners using alternative words during the original intervention when the multi-word expressions could not be quickly recalled. While trying to recall is part of the learning process, failure to recall and use of alternatives results in foregoing a chance to speak and communicate using the multi-word expressions. Therefore, in the latter interventions with the decreasing-time roleplay, multi-word expressions for the conversations were displayed on the class screen to help the learners recall and use them in their conversations. This additional opportunity to refer to the expressions would have increased the use of the expressions during the decreasing-time roleplay activity. The implication here is that if the goal of an activity is to use certain expressions, ensuring a reference point for those who are struggling will support greater use. The display of the expressions also raises the awareness of learners to try to use the expressions, rather than reverting to the path of least resistance, and using different words and expressions.

The perceptions of activity usefulness overall appeared to increase from the first cohort to the latter cohorts. The only activity that did not significantly change in terms of learner response to usefulness was phrase instruction. All others (including those where no substantial changes were made) showed significant improvements in the perception of usefulness between the first cohort and the latter cohorts. We can expect that, like many things in life, teaching improves with practice. In the latter iterations, I was more familiar with the activities, and with how the students were likely to respond based on the original intervention experience. The focus on learner feedback and teacher reflection resulted in greater improvements than would have been realised through simple repetition. The action research cycle provided structure to observe and actively seek feedback on my practice. Student feedback raised my awareness of the learner experience and thereby helped me to improve my instructions and the learning environment for the students.

In summary, the fluency workshop model using a dialogue format can improve learners' knowledge and use of multi-word expressions. The intense focus and time spent on learning and using multi-word expressions in this classroom model comes at the expense of other types of learning (such as wider vocabulary etc.). If learning goals are to improve use of common multi-word expressions and speaking confidence for particular situations then introducing this approach is likely to be beneficial. This research has provided evidence that action research can help teachers to improve the outcomes for learning multi-word expressions in the classroom. Implications for teaching multi-word expressions reported here may also apply to teaching English as a foreign or additional language more generally. Teaching is likely to improve with practice, especially if a conscious effort is made to observe and reflect on the classroom dynamic and feedback from learners. Overall, the time spent encountering and using the multi-word expressions can be maximised by reducing the time spent introducing or starting an activity. To this end, teachers may consider displaying a countdown timer and providing scaffolding assistance such as pictures to set a scenario or reference points to prompt the use of target multi-word expressions.

Of course, there are limitations which should be considered when interpreting the results of this report. As already mentioned, the original study had a five-point Likert scale while the improved studies had a six-point Likert scale, and the responses were classified as binary in order to make comparisons between the two scales. The original study participants were not forced to choose between [agree] and [disagree], and we cannot know how they would have chosen if they were. For example, there may have been more agreement or more disagreement, changing the comparison with the improved studies. Ideally, the studies should have had the same Likert scale to make the comparison more direct. Secondly, individual learning preferences may have influenced the responses in different ways compared to if the original respondents were to experience and respond to the improved activities.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance and approved by the Human Ethics Committee of VICTORIA UNIVERSITY OF WELLINGTON (Ethics approval: 22354, 24 November 2015).

Informed Consent Statement: Informed consent was obtained from all the subjects involved in the study.

Data Availability Statement: The original data presented in the study are openly available in [OSF] at https://osf.io/bvk8r/?view_only=c87b36efc09a412aa38e78762024ffd7 (accessed on 8 December 2023).

Acknowledgments: Appreciation is conveyed to all the participants who made these studies possible. Thank you to Averil Coxhead and Frank Boerswho supervised this Ph.D. research at Victoria University of Wellington. Thank you to the fellow researchers and anonymous reviewers who have asked pertinent questions and given helpful feedback along the way.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Frequency Profiles of the Multi-Word Expressions Used in the Fluency Workshop

Unit	Target Multi-Word Expression	Frequency Profile of Vocabulary	Occurrences per Million (COCA)	Function
Café	We have about an	K1	0.62	Information
Café	I usually only have	K1	0 (COCA), 1.2B (Google Search)	Information
Café	Have on the menu	Have on (K1), menu (K2)	0.05	Information
Café	What would you like	K1	3.30	Question
Café	But so does the	K1	0.07	Comparison
Café	What are you going	K1	11.25	Question
Café	I think I will	K1	0.71	Information
Café	I will have the	K1	0.13	Information
Café	Are you going to	K1	44.02	Question
Café	To get a drink	K1	0.19	Question
Directions	How do I get	K1	0.82	Question
Directions	Can you tell me	K1	5.03	Question
Directions	When you get to	K1	2.76	Instruction
Directions	Turn left at the	K1	0.03	Instruction
Directions	Until you get to	K1	0.35	Instruction
Directions	You should see the	K1	0.33	Instruction
Directions	Right-hand side of	K1	0.07	Information
Directions	Until I get to	K1	0.10	Information
Directions	Is on the right	K1	0.70	Information
Directions	Have a good day	K1	1.56	Greeting
Hotel	How many are in	K1	0.09	Question
Hotel	Could you repeat that	Could you (K1), repeat (K2)	0.15	Question
Hotel	How do you spell	How do you (K1), spell (K2)	0.14	Question
Hotel	How will you be	K1	0.12	Question
Hotel	That will be fine	K1	0.07	Information
Hotel	It is an extra	K1	0.03	Information
Hotel	Would you like to	K1	11.98	Question
Hotel	Let me think about	K1	0.22	Response
Hotel	The details are in	The, are, in (K1), details (K2)	0.03	Information
Hotel	Thank you very much.	K1	124.09	Response

References

Arevart, Supot, and Paul Nation. 1991. Fluency improvement in a second language. *RELC Journal* 22: 84–94. [CrossRef]

Boers, Frank, and Chau Thai. 2017. Repeating a monologue under increasing time pressure: A replication of thai and boers (2016). *The TESOLANZ Journal* 25: 1–11.

Boers, Frank, and Seth Lindstromberg. 2012. Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics* 32: 83–110. [CrossRef]

Boers, Frank, June Eyckmans, Jenny Kappel, Hélène Stengers, and Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10: 245–61. [CrossRef]

Boers, Frank, Thuy Bui, Julie Deconinck, Hélène Stengers, and Averil Coxhead. 2023. Helping learners develop autonomy in acquiring multiword expressions. *The Modern Language Journal* 107: 222–41. [CrossRef]

Burns, Anne. 2010. *Doing Action Research in English Language Teaching: A Guide for Practitioners*. London: Routledge.

Davies, Mark. 2013. Corpus of Contemporary American English (COCA). Available online: https://corpus.byu.edu/coca/ (accessed on 29 June 2022).

EF English Proficiency Index. 2023. Signum International AG. Available online: https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2023/ef-epi-2023-english.pdf (accessed on 9 June 2024).

Godwin, Karrie E., Howard Seltman, Ma Almeda, Mandi Davis Skerbetz, Shimin Kai, Ryan S. Baker, and Anna V. Fisher. 2021. The elusive relationship between time on-task and learning: Not simply an issue of measurement. *Educational Psychology* 41: 502–19. [CrossRef]

Jiang, Jingyi. 2009. Designing pedagogic materials to improve awareness and productive use of L2 collocations. In *Researching Collocations in Another Language: Multiple Interpretations*. Edited by Andy Barfield and Henrik Gyllstad. London: Palgrave Macmillan UK, pp. 99–113.

Koizumi, Rie, Toshie Agawa, Keiko Asano, and Yo In'nami. 2022. Skill profiles of Japanese English learners and reasons for uneven patterns. *Language Testing in Asia* 12: 53. [CrossRef]

Lin, Phoebe M. S. 2019. Formulaic language and speech prosody. In *Understanding Formulaic Language*. A Second Language Perspective. London: Routledge, pp. 78–94.

Loewen, Shawn, and Luke Plonsky. 2016. *An A–Z of Applied Linguistics Research Methods*. London: Palgrave Macmillan.

Majuddin, Elvenna, Anna Siyanova-Chanturia, and Frank Boers. 2021. Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition* 43: 985–1008. [CrossRef]

Maurice, Keith. 1983. The fluency workshop. *TESOL Newsletter* 17: 29. Available online: https://files.eric.ed.gov/fulltext/ED289347.pdf (accessed on 17 October 2024).

McGuire, Michael, and Jenifer Larson-Hall. 2018. Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal* 34: 1–25. [CrossRef]

Nation, Paul. 2007. The four strands. *Innovation in Language Learning & Teaching* 1: 2–13. [CrossRef]

Olipas, Cris Norman P., and Ruth G. Luciano. 2020. Understanding the impact of using countdown timer on the academic motivation and computer programming anxiety of IT students: The case of a state university in the Philippines. *International Journal of Scientific & Technology Research* 9: 235–40.

Peters, Ann M. 1983. *The Units of Language Acquisition*. New York: Cambridge University Press. Available online: https://www.jstor.org/stable/413467?origin=crossref (accessed on 23 March 2017).

Peters, Elke, Eva Puimège, and Paweł Szudarski. 2023. Repetition and Incidental Learning of Multiword Units: A Conceptual Multisite Replication Study of Webb, Newton, and Chang (2013). *Language Learning* 73: 1211–51. [CrossRef]

Schmidt, Richard. 2001. Attention. In *Cognition and Second Language Instruction*. Edited by Peter Robinson. Cambridge: Cambridge University Press, pp. 3–33.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11: 129–58. [CrossRef]

Siyanova-Chanturia, Anna, and Ron Martinez. 2015. The idiom principle revisited. *Applied Linguistics* 36: 549–69. [CrossRef]

Stengers, Hélène, Frank Boers, Alex Housen, and June Eyckmans. 2010. Does 'chunking' foster chunk-uptake? In *Fostering Language Teaching Efficiency Through Cognitive Linguistics*. Edited by Sabine De Knop, Frank Boers and Antoon De Rycker. Berlin: De Gruyter Mouton, pp. 99–117.

Strong, Brian, and Frank Boers. 2019. The Error in Trial and Error: Exercises on Phrasal Verbs. *TESOL Quarterly* 53: 289–319. [CrossRef]

Szcześniak, Konrad. 2024. The noticing hypothesis and formulaic language. Learnability of Non-salient language forms. *Acta Psychologica* 248: 104372. [CrossRef]

Tavakoli, Parvaneh, and Takumi Uchihara. 2020. To what extent are multiword sequences associated with oral fluency? *Language Learning* 70: 506–47. [CrossRef]

Thomson, Haidee. 2018. Building speaking fluency with multiword expressions. *TESL Canada Journal* 34: 26–53. [CrossRef]

Thomson, Haidee, Averil Coxhead, Frank Boers, and Paul Warren. 2023. Increasing use of multi-word expressions in conversation through a fluency workshop. *System* 113: 102994. [CrossRef]

Toomer, Mark, and Irina Elgort. 2019. The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning* 69: 405–39. [CrossRef]

Vilkaitė, Laura. 2016. Formulaic language is not all the same: Comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji Kalbotyra* 8: 28–54. [CrossRef]
Vygotsky, Lev Semenovich. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
Wajnryb, Ruth. 1990. *Grammar Dictation*. Oxford: Oxford University Press.
Webb, Stuart. 2007. The effects of repetition on vocabulary knowledge. *Applied Linguistics* 28: 46–65. [CrossRef]
Webb, Stuart, Jonathan M. Newton, and Anna Chang. 2013. Incidental learning of collocation. *Language Learning* 63: 91–120. [CrossRef]
Wood, David. 1998. *Making the Grade: An Interactive Course in English for Academic Purposes*. Toronto: Prentice Hall Allyn and Bacon.
Wood, David. 2009. Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics* 12: 39–57.
Wray, Alison. 2004. 'Here's one i prepared earlier': Formulaic language learning on television. In *Formulaic Sequences Acquisition, Processing, and Use*. Edited by Norbert Schmitt. Amsterdam and Philadelphia: J. Benjamins, pp. 249–68.
Wray, Alison, and Tess Fitzpatrick. 2010. Pushing learners to the extreme: The artificial use of prefabricated material in conversation. *Innovation in Language Learning and Teaching* 4: 37–51. [CrossRef]
Wulff, Stefanie. 2019. Acquisition of formulaic language from a usage-based perspective. In *Understanding Formulaic Language: A Second Language Acquisition Perspective*. Edited by Anna Siyanova-Chanturia and Ana Pellicer-Sanchez. New York: Routledge, pp. 19–37.
Wurtele, Sandy K., and Ronald S. Drabman. 1984. 'Beat the buzzer' for classroom dawdling: A one-year trial. *Behavior Therapy* 15: 403–9. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Retelling of Stories with Common Phrasal Expressions by High-Proficiency Learners: Implications for Learning and High-Stakes Testing

David Gregory Coulson

Graduate School of Language Education and Information, Ritsumeikan University, Kyoto 603-8577, Japan; coulson@fc.ritsumei.ac.jp

Abstract: The goal of this research was to investigate how well L2 English speakers of different proficiency levels were able to perform on a test of auditory memory like that used in the Wechsler Memory Scale (WMS). In this test, participants must retell heard stories. While the validity of elicited imitation tests is well established in second-language acquisition research, the effectiveness of using retelling as a neuropsychological measurement when a language other than the test taker's L1 is used is unclear. Further, due to their importance in memory function, this study also analyzed the role of common phrasal expressions in how well participants retold stories in three groups of different proficiency. The results indicated that the increase in scores in the retelling of stories aligned with the proficiency level of the non-native participants. Higher-proficiency NNSs were also able to use more of the commonly used spoken phrases in their retelling. Further, there was no difference in this measure between the higher-proficiency Second-Language English and First-Language English participants. While the effectiveness of this test method as a neuropsychological tool when a language other than the test taker's L1 is used is unclear, these results indicate significant progress by the higher-proficiency participants. Nevertheless, given that this test is often used in this context with immigrants and minority language patients, doubts remain about its suitability for less proficient individuals.

Keywords: elicited imitation; memory ability; lexical knowledge; second-language acquisition

1. Introduction

Working memory (WM) is an essential mechanism for language use that allows humans to process briefly stored information. Through focused attention, it is possible to manipulate information necessary for the completion of particular tasks and to suppress that which is not relevant to the task in hand. According to Haarman (2013), WM can be used with "tasks that require immediate recall of a list of items, either in the form of a simple span or complex span task" (p. 696). This can be assessed through both reading and listening, during which participants are presented with information and are then asked to recall it. The ability to recall information is helped by the presence of prefabricated chunks of language that facilitate the production and recall of information. However, as Siyanova-Chanturia and Janssen (2018, p. 2010) explained, the benefit of these chunks is much greater for native speakers than non-native speakers, for whom phrases are frequently unfamiliar. Formulaic phrases are generally processed faster than novel phrases by native speakers. These help speakers to bypass short-term memory, and this leads to faster, more automatic processing. This characteristic of first-language skill is due to the massive amount of input that is available to people who have grown up with a particular language. This also implies that non-native speakers are at a disadvantage in the processing of these phrases.

As a widely used test of neuropsychological function, The Wechsler Memory Scale (WMS) measures various aspects of both short-term, long-term and working memory ability. Subtests include auditory, visual, immediate and delayed memory. One of the

tasks, called the Logical Memory Subtest (LMS), involves story retelling during which a clinician reads short stories to the patient, who is then asked to retell the story immediately. These are significantly longer than texts used in Elicited Imitation (EI) tests that have been widely researched in second-language learning and found to be reliable, valid measures of second-language proficiency (e.g., McManus and Liu 2020).

Phonological short-term memory (PSTM) refers to the ability to temporarily store and recall speech sounds for a short period of time (Baddeley 2015). This cognitive function is essential for the manipulation of auditory information involved in the understanding and production of language. Usually conducted in a speaker's L1, the Wechsler Memory Scale (WMS) can be used to evaluate PSTM by having participants listen to and verbally recall stories told during the test (Gallagher and Azuma 2018). This provides insights into an individual's memory retention and linguistic capabilities, which are crucial for diagnosing cognitive impairments, language disorders and other neuropsychological conditions.

Research on second-language acquisition has shown that PSTM also plays a significant role in L2 acquisition, supporting not only vocabulary acquisition but also general proficiency. For example, Kormos and Sáfár (2008) suggested that the ability to access and encode vocabulary quickly, and to fluently form lexical chunks, is related to phonological short-term memory. During the listening to, recalling and retelling of stories, L1 speakers have distinct advantages over L2 speakers. This is because listening relies much more on top-down processing than reading (Qian and Lin 2020) due to its cursory nature. Understanding of spoken language requires the activation of short-term working memory, and L1 speakers have a much more strongly developed ability of a language. Further, L1 speakers can draw upon a more extensive, and deeply entrenched, lexical and syntactic knowledge base, which allows them to encode and retrieve information more efficiently. They are also typically more familiar with the cultural and contextual nuances embedded in stories, aiding their recall and retelling abilities. Importantly, due to their large and strongly automatized phrasal knowledge, L1 speakers devote fewer cognitive resources to processing language. There is a clear advantage from all kinds of formulaic language, compared to non-formulaic language, to the speed of understanding (e.g., Carrol and Conklin 2020), freeing up more working memory (the fundamental cognitive apparatus for the storage and processing of information) which can be used to recall the details of a story. In contrast, L2 speakers often need to allocate more cognitive resources to understand and produce the language itself, leaving less available for recalling story details and structuring a coherent retelling.

This study examines the effects that the factors described above may have on L2 users at different proficiency levels. This is important since it is not uncommon for cognitive tests such as the WMS to be given to immigrants or minority language patients in their L2. Therefore, this study examines how the ability to use multiword expressions (MWEs) assists in the retelling of stories used in this kind of assessment on different groups of participants, including L1, medium- and high-proficiency L2 speakers.

2. Literature Review

2.1. Memory and Formulaic Language in Native Speakers and Second-Language Users

Research consistently shows that a substantial portion of language is composed of MWEs that frequently co-occur in both spoken and written texts. For example, a review of relevant research on the proportion of formulaic phrases in spoken and written discourse by Conklin and Schmitt (2012) indicated that they constitute up to about half of such texts.

The processing of both words and phrases in memory is essential to language use. However, non-native speakers are often at a disadvantage in the amount of exposure to a second language they receive, whereas monolinguals are able to process language faster thanks to far richer exposure which results in "lexical entrenchment" (Conklin and Thul 2023, p. 208). In particular, during listening and reading, MWEs are processed faster and to a greater degree, the higher their frequency is. Conversely, non-native speakers often do not process MWEs as units, mainly due to a lack of encounters. It is reported that

even highly proficient learners do not derive as large a benefit from formulaic language as L1 speakers. For example, research by Lundell and Lindqvist (2013) with advanced French L2 learners examined whether lexical ability continues to develop after reaching a very high level of proficiency. Although their Swedish participants had lived, on average, 14.5 years in France, the results of their study showed no significant correlation between lexical ability and the length of residence, although several L2 subjects had reached a similar score as the NS subjects. Similar results are also found in more recent research. The results of a meta-analysis on the processing advantage of multiword expressions by Yi and Zhong (2023) also supported the centrality of formulaic language. Both L2 as well as L1 speakers derived a lot of benefit from the processability of such language. The authors stated that in certain tasks, including natural reading, "L2 speakers may also engage in implicit processing of MWSs" (p. 19), although the advantage was slightly greater for L1 speakers. Nevertheless, less proficient L2 learners usually process formulaic language on a word-by-word basis.

A significant benefit of formulaic language is that it can increase the amount of linguistic information that a speaker can deal with. Yang (2015) described how advanced Chinese L1 speakers of English were able to recall more words than mid-proficiency participants in a reading span task. Nevertheless, L2 speakers do not enjoy as strong an implicit lexical ability as L1 speakers (e.g., Conklin and Schmitt 2012).

Research by Sonbul (2015) showed a significant difference in the speed of processing of higher and lower-frequency collocations both in L1 and high-proficiency L2 speakers. Schmitt et al. (2004) compared the ability of L1 speakers, L2 post-graduate students and overseas scholars (with very advanced English proficiency). The test involved the participants listening to stretches of dictations which were beyond regular short-term memory span. Even some of the First-Language English (FLE) participants were unable to express the content of the texts with all the formulaic phrases intact. However, their ability was beyond that of the Second-Language English (SLE) participants. The latter participants recreated the texts by focusing on key content words, which was probably due to their lack of knowledge of recurrent clusters that had been derived from a corpus for the study.

Finally, and of particular relevance to the focus of this study, research by Kisser et al. (2012) examined the performance of FLE and advanced SLE participants on various neuropsychological tests. All were university students in America. Their results showed that while the results of non-language dependent tasks were at a similar level, there was a small but statistically significant gap between the two groups on tests involving language processing.

Taken together, the role of MWEs, including phrases, lexical bundles and collocations, can be clearly understood as facilitative of language processing, assisting with the understanding and production of language. Nevertheless, in second-language acquisition, a persistent gap often remains between L1 speakers and even very advanced L2 language users in receptive and productive skill associated with MWEs.

2.2. The Phrasal Expressions List

Collocations are "notoriously difficult for learners" according to Paquot and Granger (2012, p. 130) and "pose considerable difficulties, even for the advanced learner" (Nesselhauf 2005, p. 2). Studies have shown that L2 learners are just as likely to produce incorrectly collocated bundles as they are to produce correct collocations because there is a tendency on the part of L2 users to focus on the key words of the collocations rather than the whole formulaic expression (Hoang and Boers 2016). Because of this, it can be challenging to decide what constitutes formulaic language. In order to create a useful test of implicit collocational skill, it is crucial to select phrases that have been carefully validated. To this end, the Phrasal Expressions List (PHRASE list), created by Martinez and Schmitt (2012), is a highly valuable resource. It contains 505 MWEs that are used in spoken, written and academic English. The phrases mainly consist of words from the most frequent 2000 words. At least one of the following criteria for the inclusion of phrases in the list had

to be met: (a) whether they were Morpheme Equivalent Units, meaning a phrase that is processed as one unit with a single meaning; (b) whether the expressions were semantically transparent; (c) whether they were 'deceptively transparent' (p. 308). In particular, the authors considered that it would not be useful to include phrases whose meanings can be easily guessed from their surface form. Rather, more opaque phrases such as *every so often* (which might be misunderstood by learners as meaning *often* whereas it actually means *sometimes*) were included.

In the PHRASE list, the typical discourse genre in which each of the 505 phrases is used is shown. These are either Spoken general, Written general or Written academic. Table 1 shows two examples from the list: the most common phrase (appearing over 83 thousand times per 100 million words) is *have to*. It is commonly used in Spoken general discourse, as indicated by the three stars (***) in that column. The phrase *as long as* is used commonly in both spoken and written genres, as also shown by the three stars in those columns. However, neither of these expressions are used commonly in written academic texts, as indicated by the single star (*) in that column.

Table 1. Examples from Martinez and Schmitt's PHRASE List.

Phrase	Frequency (per 100 Million)	Spoken General	Written General	Written Academic	Example
HAVE TO	83902	* * *	* *	*	I exercise because I **have to**.
AS LONG AS	5084	* * *	* * *	*	It makes no difference **as long as** it's done.

As described below, for this research, phrases with a three-star rating on the Spoken general category were selected for use in the creation of the short stories for the EI retelling task. Phrases from the Spoken category were considered to be the most appropriate because the stories would be given orally and also because these are the phrases the L2 learners would be more likely to acquire in their daily lives in New Zealand.

2.3. Testing of Implicit Knowledge of Second Languages

Elicited Imitation (EI) is a testing method in which participants listen to a stimulus and are then instructed to repeat it as accurately as possible (Kostromitina and Plonsky 2022). It has been used widely as a methodology for assessing the implicit ability of second-language learners and users, particularly in the area of grammatical accuracy. According to Solon et al. (2019, p. 1030), the demands placed on non-native speakers by EI are greater than those on native speakers. The authors stated that native speakers can rely on automatic processing during EI, whereas the burden on the working memory capacity of SLEs may be higher. Kostromitina and Plonsky (2022) reported in their meta-analysis that the longer the sentences were, the greater the cognitive load for participants became. As such, their reliance on working memory decreased, and they had to rely more on their linguistic knowledge. Further, Solon and Park (2024) pointed out that while elicited imitation has a solid research background, more validation is necessary to investigate its complex nature. Particularly concerning the focus of this study, more research is necessary to clarify whether EI is as equally effective for learner populations beyond the usual focus on higher-proficiency participants. Concerning this, in research by Stålhammar et al. (2022), the ability of both native and non-native, but fluent, Swedish speakers was assessed using a neuropsychological test. It was found that the native speakers performed significantly better than the non-native participants on subtests with a verbal component. These individuals were from 34 different countries and so represented a wide range of language backgrounds. This is important as participants whose first language was further from Swedish scored lower (p. 832).

Research, such as that by Deygers (2020), has shown that native speakers are better able to achieve high scores on EI tests than non-native speakers, due to their solid command of the language that the EI test is given in. In particular, working memory span in humans is not long, and if the length of the test item exceeds this, testees have to rely more on using implicit linguistic knowledge to provide as accurate a response as possible. As Deygers reported (p.934), "the rationale for the use of EI as a measure of language proficiency relies on the consensus within the field of neuropsychology that working memory cannot hold on to aural information for very long". This is relevant to the present study as there is a clear overlap between the EI methodology and the LMS test used in the WMS.

3. Background to the Current Study

It was the experience of the author to need to be clinically assessed on the Japanese version of the WMS. I was living in Japan at the time, and the assessment was only available to be conducted in Japanese, my third language, albeit one that I had been using for over 25 years.

As a non-native speaker of Japanese, the subtest which caused the most consternation was the LMS in which verbatim recall of several stories, one by one, is required. It became clear while trying to repeat the content of these stories that the clinician was discretely marking the number of correctly recalled parts ("idea units" as termed below) of the story. The stories correspond to a length of about 50 words in English. Therefore, this can be considered as prolonged form of the EI test format.

As seen in the review of related literature, this retelling approach is very challenging for non-native speakers, even for those with advanced command of the language being used. Upon receiving an initial explanation of the test content from the clinician, I expressed concern that having to recall the content in my third language was unlikely to result in a reliable assessment of my memory ability. While this concern about the validity of the WMS assessment was acknowledged, there was no alternative but to submit to this assessment as this was a Japanese-medium clinical environment.

As such, the clinical practice of assessing non-native speakers in their second language for memory skill, even for those of high proficiency, raises concerns about the validity of this approach. Concerning this, the main goal of this paper was to conduct an investigation into how well FLEs perform on such an assessment compared with SLEs. This issue is particularly relevant in an era of large-scale immigration in countries across the globe, which is only likely to increase in the near future.

In particular, as the focus of this assessment was to investigate short-term memory ability during retelling, it is clear that the role of MWEs is a very pertinent variable. As such, the research questions are:

1. How do SLE participants of different proficiency levels score on retelling stories of increasing length, compared to FLE participants?
2. How do high-frequency phrases contribute to the ability to retell the four stories of increasing length?
3. What is the contribution of the number of phrases used in each story to the score, and how does this relate to the proficiency level of the participants?

4. Participants

The SLE participants in this research were 42 L1 speakers of Mandarin who were enrolled as students in a university in New Zealand. There were two groups. Group 1 was composed of 20 undergraduate students with either IELTS 5.5 or 6.0 proficiency (referred to as 'modest' to 'competent' ability in IELTS bands). Their mean length of residence in New Zealand was 0.63 years (SD 1.19). Group 2 was composed of 22 mainly post-graduate students with IELTS 6.5 and above ('competent' to 'good' in IELTS). Their mean length of residence was 1.55 years (SD 1.34). Group 3 was composed of 12 FLEs.

5. Methodology

Four short stories were created for the story-retelling task (Appendix A). They were all 52 or 53 words in length. To investigate the retelling ability of participants at different levels of proficiency, the length of the EI sections increased across the four stories. Finally, story 4 approximately matched the length of the story retelling text used in the LMS.

Further, to investigate the contribution of phrasal knowledge to the recall of the stories, six MWEs from the PHRASE list were included in each of the four stories. To build confidence, and hopefully not overwhelm the participants, the first two stories were told in short sections, five in story 1 and four in story 2. The first story (five sections) is shown below. The pauses between the EI sections are marked by //, and the six phrases are highlighted in grey.

Story 1

Last year, John started university. // His mother said he should make meals on his own . // To make sure he did this, she taught him some recipes. // He managed to continue for a few weeks, but he soon gave this up . // In the end , he only ate high-calorie dishes in the campus cafeteria.

(52 words)

Story 3 (Medium) was told in two sections of 27 and 25 words. Finally, story 4 (53 words in length) was told with no pause. It is this fourth story which is closest to the conditions under which the story recall task in the LMS is conducted.

The six phrases in each story were selected from the 'Spoken general' column (each with three stars) in the PHRASE list. The average frequency of the phrases is 4939 per 100 million words. In addition, the four EI stories were written in simple English, with 88% of the words in the story coming from the first 1000-word band of the COCA list. The aim of this design was to strengthen the focus on the ability to retell the content and to assess the ability of the participants to productively recall the six phrases in each story.

The texts used in the four EI retelling tests for this research were written based on the general characteristics of the content of the Logical Memory Subtest. This is part of the WMS and is used in the assessment of memory disorders and cognitive impairment. Participants are asked to listen to short stories and then recall the content in as much detail as possible, both immediately and after a delay. The purpose is to assess attention, comprehension and working memory. The clinician briefly explains the procedure and simply asks that as much of the content as possible be repeated. Having taken the same LMS several times, I was able to construct similar texts with specific characteristics, described above, that were necessary for this research. It is important to state that these texts are completely new and do not infringe upon the copyright of the original tests.

The tests were conducted one-to-one in a quiet room. Participants' responses to the EI tested were recorded. The participants were fully informed about the nature and purpose of the research. They all signed a consent form and were paid the equivalent of six US dollars for their time. Victoria University of Wellington, where the testing was conducted, gave the research procedure ethical approval. Short story 1 was read first and the participants' recall was recorded. After breaks of about 20 seconds, stories 2, 3 and 4 were read. Participants' recorded responses were subsequently checked carefully.

The test answers were assessed in the following way: the total number of "idea units" that could be repeated by the participants from the four texts was counted. Here, "idea units" refers to the discrete ideas contained in each of the texts to be used for the retelling task. The total maximum score for the four stories was 100 (Appendix B). In addition, the number of MWEs from the PHRASE list that were accurately recalled during the retelling was counted, with a maximum score of 24 (six in each story).

In summary, the retelling test developed for this study consists of four stories which were read out in sections of increasing length until the final uninterrupted whole story of 53 words was read. The participants were assessed on two measures: their immediate

recall of the 100 idea units, and their accurate use of the 24 phrases while retelling the four stories.

6. Analysis and Results

Concerning the first research question (How do SLE participants of different proficiency score on retelling stories of increasing length, compared to FLE participants?), the results for the three groups (moderate, competent and FLE) on the three story lengths are shown in Figure 1. It can be seen that the scores for the moderate group fell most sharply after the short story, whereas the competent group showed a less pronounced decline. The scores for the competent and FLE groups showed a complex pattern. On the medium-length text (Story 3), the competent group scored higher than the FLE group. This did not occur for the long story 4. A more detailed analysis on this is provided below.

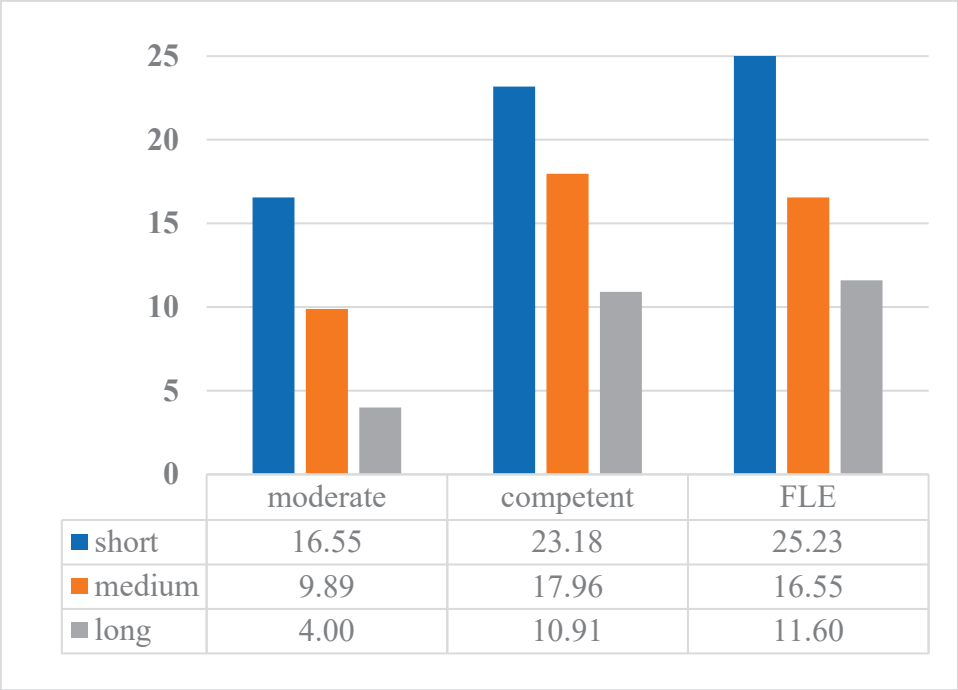

Figure 1. Scores on idea units from retelling at different story lengths.

Multivariate Analysis of Variance (MANOVA) was conducted to examine the effects of proficiency levels and story length on students' scores. There was a significant main effect for proficiency level, $F(2, 235) = 120.681$, $p < 0.001$, and a significant main effect for story length, $F(2, 235) = 323.824$, $p < 0.001$, on the total scores. However, the interaction between proficiency level and story length was not statistically significant, $F(4, 235) = 1.968$, $p = 0.100$. This indicates that while both proficiency level and story length independently affected students' scores, the effect of story length on scores did not significantly differ across proficiency levels. That is, the change in total scores with increasing proficiency was consistent regardless of whether the story was short, medium or long.

Post hoc comparison tests were conducted with a Bonferroni correction to compare the mean differences in total scores across proficiency levels for each story length. The post hoc tests revealed that total scores significantly improved for all story lengths as proficiency levels increased: IELTS 5.5 and 6, IELTS 6 and above, and the FLE level. A significant improvement in total scores was observed as the proficiency level increased from 6.5 and above to FLE level for short story lengths, but no such significant improvement was recorded for the medium and long story lengths (Table 2).

Table 2. Total score across different proficiency levels.

	Mean Total Score (SE)			Changes in Total Score		
	Proficiency Level					
Story Length	**IELTS Modest**	**IELTS Competent**	**FLE**	**5.5~6 and 6.5 and Above**	**6.5 and Above to FLE**	**5.5~6 to FLE**
Short	16.55 (0.53)	23.18 (0.49)	25.23 (0.52)	6.629 **	2.043 *	8.672 **
Medium	9.89 (0.75)	17.96 (0.70)	16.55 (0.73)	8.060 **	−1.405	6.655 **
Long	4.00 (0.75)	10.91 (0.70)	11.60 (0.73)	6.909 **	0.691	7.600 **

Note: Numbers in parentheses are standard errors. * $p < 0.05$ (Bonferroni correction) and ** $p < 0.001$ (Bonferroni correction).

Concerning the second research question (How do high-frequency phrases contribute to the ability to retell the four stories of increasing length?), a post hoc comparison test with a Bonferroni correction was conducted to examine whether the total number of phrases that participants used significantly increased as the proficiency level increased. The results showed that the number of phrases used increased significantly at the 0.01 level for all story lengths between the two IELTS groups (5.5 and 6 and 6.5 and above), and the IELTS 5.5 and 6 group and the FLE group. However, there was no significant increase in the number of phrases used by students between the 6.5 and above, and FLE proficiencies across all story lengths (see Table 3).

Table 3. Numbers of phrases used across different proficiency levels.

	Mean Number of Phrases (SE)			Changes in Number of Phrases		
	Proficiency Level					
Story Length	**Modest**	**Competent**	**FLE**	**Modest to Competent**	**Competent to FLE**	**Modest to FLE**
Short	4.26 (0.18)	5.64 (0.16)	6.00 (0.17)	1.373 **	0.364	1.737 **
Medium	1.58 (0.25)	3.41 (0.23)	3.65 (0.24)	1.830 **	0.241	2.071 **
Long	0.58 (0.25)	3.14 (0.23)	2.80 (0.24)	2.557 **	−0.336	2.221 **

Note: Numbers in parentheses are standard errors. * $p < 0.05$ (Bonferroni correction) and ** $p < 0.001$ (Bonferroni correction).

These results indicate that as proficiency level increased, students tended to use a greater number of phrases. This trend was consistent across all story lengths when comparing the IELTS groups 5.5 and 6 and 6.5 and above, and the IELTS 5.5 and 6 group and the FLE group. The fact that the IELTS competent group actually repeated more phrases in the long story than the FLE group (3.14 versus 2.80) suggests that while the FLE group may have been able to recall the events of the story better than the IELTS competent participants, this may have been due to the need to allocate fewer cognitive resources to process the linguistic features of the story. As a result, they may have been better able to recall the story correctly without the use of MWEs.

Concerning the third research question (What is the contribution of the number of phrases used in each story to the score, and how does this relate to the proficiency level of the participants?), in order to analyze the contribution of the number of phrases used in each story to the total idea unit score, a Linear Regression analysis was conducted utilizing Hayes' Process v3.5. The aim of this analysis was to investigate the relationship between the number of phrases used in each story and the total idea unit score attained by the participants and to examine whether the proficiency level moderated this relationship.

For the first short story, the number of phrases explained 89.9% of the variability in scores ($R^2 = 0.899$, $F(1,59) = 522.50$, $p < 0.001$). When both the number of phrases and proficiency level were included as predictors, they explained 93.6% of the variability ($R^2 = 0.936$, $F(3,57) = 279.75$, $p < 0.001$). Similarly, for the second short story, the num-

ber of phrases explained 60.5% of the variability in scores (R^2 = 0.605, F(1,59) = 90.448, p < 0.001), and the combined model with both predictors explained 72.3% of the variability (R^2 = 0.723, F(3,57) = 49.557, p < 0.001). However, the interaction term between the number of phrases and the proficiency level was not significant for either of these stories, indicating that proficiency level did not significantly moderate the contribution of the number of phrases to the final score that participants received for retelling the four stories.

For the medium story, the regression analysis showed that the number of phrases accounted for 59.3% of the variability in scores (R^2 = 0.593, F(1,59) = 86.06, p < 0.001). Including proficiency level as a predictor increased the explained variability to 78.2% (R^2 = 0.782, F(5,55) = 39.520, p < 0.001). The interaction term was significant (ΔR2 = 0.0877, ΔF(2,55) = 11.073, p = 0.0001), indicating that proficiency level significantly moderated the contribution of the number of phrases to the score for the medium-length story. The regression analysis showed that the number of phrases explained 67.0% of the variability in scores (R^2 = 0.670, F(1,59) = 119.89, p < 0.001). The combined model with both predictors accounted for 79.4% of the variability (R^2 = 0.794, F(5,55) = 42.51, p < 0.001). Similar to the medium-length story, the interaction term for the long story was significant (ΔR2 = 0.0826, ΔF(2,55) = 11.05, p = 0.0001), showing that the level of proficiency significantly moderated the relationship between the number of phrases and the score for the long story.

Overall, the regression analyses indicated that the number of phrases used in the retelling of the stories significantly contributed to the total scores across all story lengths (see Figure 2). The moderation effect of proficiency level was not significant for the short stories but was significant for the medium and long stories. This suggests that higher proficiency learners were better able to use phrases effectively in more complex storytelling contexts, leading to higher scores. The number of recalled phrases was lowest for the long story, and this indicates how the highest-scoring participants also did not repeat all phrases verbatim. Conversely, on the short stories, accurate recall of the phrases was highest, almost certainly due to the lower burden on working memory.

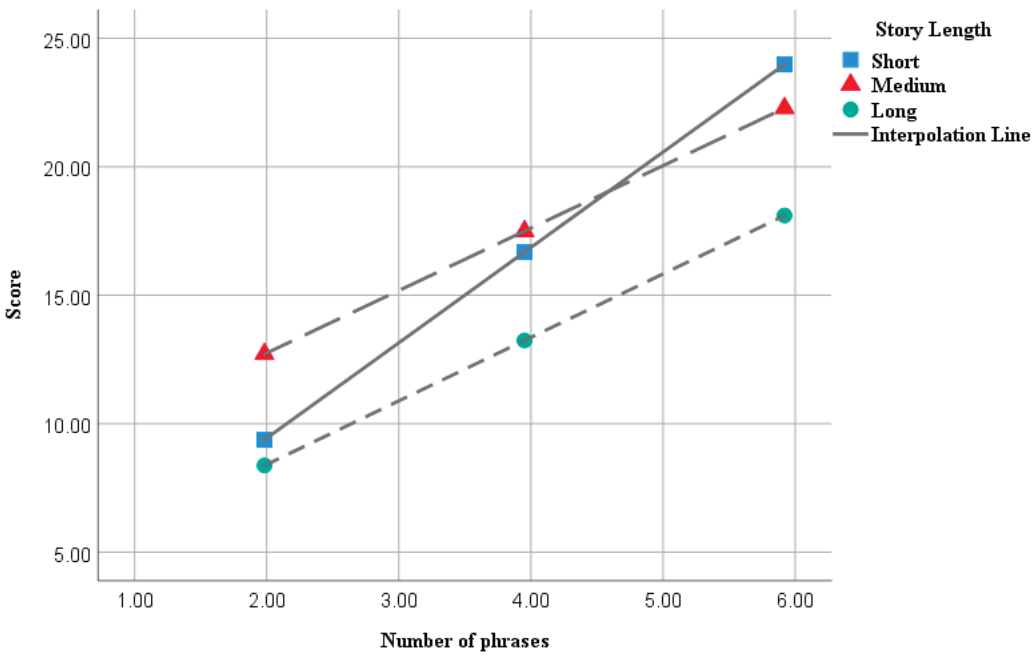

Figure 2. Relationship between the number of phrases and scores for different story lengths.

7. Discussion

Irrespective of the length of the four stories in this experiment, higher scores were seen on the recall tests across the three groups in this research. The FLE participants scored higher and recalled more of the phrasal expressions used in the elicited texts. Concerning the SLE participants, the score on idea units of the moderate group on long story 4 was

significantly less than that of the higher-proficiency competent group (4.00 vs. 10.91). This clearly shows that in a neuropsychological test such as the Logical Memory Test, there can be very significant gaps between what non-native speakers of varying second-language ability may be capable of. This is despite the fact that all the participants in this research were capable enough in English to be studying in an overseas university. This is a crucial finding that should be considered carefully in clinical assessment circles. Conversely, unexpected results also appeared. Particularly, the gap between the higher-proficiency competent group and the FLE group on the long recall test (shown in Figure 1) was only 0.69 points (10.91 and 11.60) and was very noticeable.

In sum, unless an individual has very well-developed ability in a second language, including highly developed receptive and productive collocational skills, assessment by clinical tests such as the WMS needs to be carefully examined more to ensure that the complex interaction between explicit knowledge and implicit skill can be reliably interpreted.

Concerning the gap in proficiency between the two SLE groups in this study, their average length of residence in New Zealand had not been long. The modest proficiency group had stayed 0.63 years, while the competent proficiency group had stayed 1.54 years. This signifies that while medium-term residence in a new country may sometimes indicate advanced proficiency, the reality may be much more complex. Similarly, some long-term immigrants never achieve such a high level of proficiency in the language of the country they live in despite living there for many years. In this way, clinicians should not make assumptions of memory skill based merely on length of residence.

Concerning the methodology employed in this study, the use of elicited imitation (listening to, and then recall of, four stories: two in short sections, one in the medium section and one without pauses) is supported by previous similar research. Although somewhat different from the present study, Janssen and Barber (2012) had their Spanish first-language participants recall various kinds of n-grams, and it was found that the speed of production was faster for the higher-frequency phrases. This indicates a processing advantage for more common phrases over less frequent ones. The importance of high-frequency MWEs, such as those in the PHRASE list, is clearly re-emphasized by such research. Siyanova-Chanturia and Janssen (2018) stated that more research is needed to ascertain whether this finding was not unusual. The present study has tried to address this issue, and the results align with their findings.

Further, this study has also provided grounds for believing that the use of the Logical Memory Subtest (LMS) from the Wechsler Memory Scale for the assessment of non-native speakers may not be invalid. This was shown by the somewhat unexpected ability of the competent group in this test to have scored higher than the FLE group on the medium-length story. This is impressive since the participants were speakers of Mandarin Chinese, which is a distinct language, grammatically and lexically, from English. However, the use of the LMS for neuropsychological assessment undoubtedly demands careful consideration of language proficiency. Within clinical assessment practice, such tests should be adapted to match the language and cultural background of the testees. In this way, test makers and clinicians should be aware of the potential biases present in these testing tools, and greater awareness should be paid to how to interpret the results of these tests to provide a more accurate and reliable evaluation of memory function.

Research reviewed in this paper (e.g., Stålhammar et al. (2022)) indicated a persistent gap between native- and non-native speakers on neuropsychological tests with a verbal component. Although speakers of languages further from the target language Swedish scored lower, the results of this study do not align with their findings in that Chinese L1 speakers started to come close to the "idea units" scores of the FLE participants despite the gap between these languages. Similarly, the research reported above by Lundell and Lindqvist (2013) showed that despite having lived in France for many years, no correlation between length of residence and lexical ability was found. The results of this study do not refute this, but although the SLE participants had lived only around 1.5 years in New Zealand, it was members of this group who started to approach the idea unit scores of the

FLE participants, and actually surpassed them, as seen in Table 2, with their answers to the medium Story 3 (17.96 vs. 16.55).

An important limitation of this study is that, as an exploratory task design, the EI stories were written with intentionally simple English, both in the general text and the choice of MWEs. It would be very useful to conduct the experiment again with more advanced texts. Further, if texts composed of less frequent vocabulary were used, the hypothesis that a wider divergence between the FLEs and even advanced non-native participants could be investigated.

In addition, the present study also has significance beyond the goal of neuropsychological assessment. While conducting the study, feedback from the Mandarin native-speaker participants was informally gathered. As one significant insight, concerning the need to recall heard information in an academic setting, one participant commented,

> I meet the same problem when I take part in course (sic). I can't remember clearly. I just remember some words and I don't remember the whole sentence.

It is clearly highly challenging for many students involved in academic study in their second or other language to keep in memory essential details and concepts during both listening and reading. In this way, it is essential to assess, without any implications for their grades, the ability of second-language users to function well in foreign-language study environments. As the findings by Siyanova-Chanturia and Janssen (2018) showed, how FLEs and SLEs can use phrasal language needs much more investigation. Future research could explore this issue, possibly through a similar elicited imitation task as used in this study.

Concerning implications for second-language learning, the results of this study indicate the need to reconsider the rate at which lexical skills of FLEs are thought to develop. In the EI method used in this study, points are allocated for recalled idea units rather than verbatim repetition of the language used. As reviewed above, while Kostromitina and Plonsky (2022) stated that longer texts place a heavier burden on test takers, these do illustrate the implicit ability of second-language learners. Indeed, in the long story, the competent SLE group used more of the phrasal expressions than the FLE group. These results suggest that the SLE participants had developed a high awareness of common phrases in English, and this enabled a more accurate recall of them. However, this observation cannot be taken as evidence that the SLEs' English proficiency approximated that of the FLEs, who more frequently re-expressed the text meanings in summarized form. Rather, it likely reflects a tendency among the more proficient SLE group to employ greater attention to the lexical forms of the texts. Indeed, research by Lei and Yan (2022) showed this to be the case. They investigated the kinds of strategies that learners of Chinese and L1 speakers of Chinese used on EI tasks. Through interviews, it was found that higher-proficiency learners used far more strategies than native speakers and also lower-proficiency participants. They used cognitive strategies such as paying attention to the meaning of words, phrases or syntactic patterns in the texts, or the use of translation. However, the degree of variation in the EI scores was accounted for more by factors such as the length of the items on the test rather than the use of strategies. Nevertheless, the authors stated that greater use of strategies would lead to higher scores on EI tests. Prior to participation in this study, the SLE students benefited from living and studying in English-medium academic institutions (albeit only for up to a year and a half), and their scores indicated that their productive recall of common phrasal expressions improved as the length of the stories increased. Their proficiency was high, but in addition, investigating what forms of attention they paid to learning would be very useful.

There are important implications for language education from this study. The importance of MWEs needs to be more widely acknowledged by teachers and teacher trainers as these results show that higher proficiency coincides with the acquisition of common, useful phrases. While MWEs are essential in any language, the acquisition of such phrases is nevertheless slow and incremental. MWEs include various subtypes (Wolter 2020), some of which are less important at early stages of learning, and without good awareness of

which to focus on, they tend to be underemphasized in second-language teaching. Also, the sheer number of such phrases without adequate recycling slows learning. Therefore, it is essential that language syllabi provide learners with exposure to MWEs, both to deepen learners' knowledge and build awareness for how to go on learning them strategically. Peters (2014) found that learning activities that provided repeated encounters with MWEs allowed learners to acquire more of them, although the acquisition of words was stronger than MWEs in her study.

As such, a suitable approach to syllabus design is a combination of reading activities that promote incidental learning along with explicit instruction. A well-regarded approach is Nation's Four Strands approach (e.g., Nation and Yamamoto 2012), in which learning activities can be characterized as either meaning-focused input, meaning focused output, language focused learning, and fluency development, or a combination of these four. This means that just learning through one method is inadequate to help learners acquire words. Nation and Yamamoto stress how learning occurs is more important than how teaching is done. As such, intentional learning of words should not use more than 25% of the time available for learning collocations, in this case. Rather, reading, listening and writing activities that support the development of lexical knowledge should be prioritized. In this way, Extensive Reading is a key strand in meaning focused input.

Other innovative teaching methods are also possible. Although the six phrases used in each of the four EI stories in this study are high frequency, the entrenchment in memory of such MWEs may remain shallow until learners become highly competent. To address this pedagogically, teachers can employ classroom activities such as dictogloss, in which both input through listening, output through writing and group discussion and feedback from the teacher are available. These fulfill at least three of the four strands. One such example of a pedagogic innovation is the modified dictogloss activity (Lindstromberg et al. 2016), whereby learners are given a list showing the MWE phrases to appear in the text before it is read out by the teacher, and the learners recreate the text. In doing so, the authors showed that learners were better able to use these phrases than a control group in posttests. Based on this approach, teachers could create dictogloss texts containing the most frequent MWEs in the PHRASE list, like those used in this study, and which were shown to become productively available to learners in the EI task. This could be subsequently tested with longer forms of elicited imitation than usually employed to investigate whether such important, useful phrases become productively available to learners earlier than usual.

8. Conclusions

Finally, while collocations and phrasal knowledge are central to language ability and second-language learning, this research has shown that the ability to recall heard phrasal expressions (even the comparatively common ones used in this research) is not assured. Clearly, acquisition of a second language and its lexical richness, among other features, requires a very large amount of exposure to the target language, as Conklin and Schmitt (2012) pointed out. The learning practices best suited to develop rich knowledge of a second language need greater emphasis in both schools and teacher training courses. In addition to these, there is also the need for a wider understanding of how second-language acquisition functions, and in this research, how second-language ability can influence potential outcomes in clinical neuropsychology assessments was investigated. It is hoped that this research will lead to further investigations in this area.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by Victoria University of Wellington Standing Committee of the Human Ethics Committee (approval Code 0000028128, approved 26 November 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The four stories used for the Elicited Imitation test are shown below. The six phrases in each story are highlighted in grey.

Story (1) Short Told in five short sections (// shows the sections)

Last year, John started university. // His mother said he should make meals on his own. // To make sure he did this, she taught him some recipes. // He managed to continue for a few weeks, but he soon gave this up. // In the end, he only ate high-calorie dishes in the campus cafeteria.

(52 words; Total score 25)

Story (2) Short Told in four short sections (// shows the sections)

There used to be a cake shop in my town. // The lady in charge always came up with new cakes. // No matter the weather, I went there to buy the new cake every weekend. // I kept on doing this until I left home, and I'll remember it as long as I live.

(52 words; Total score 26)

Story (3) Medium Told in two halves

British people are fond of keeping pets. They get a sense of well-being from taking care of them. They say that pets don't take them for granted. // Dogs especially can tell when their owners are sad. So, it's no wonder that some people even sleep in the same bed with their pets.

(52 words; Total score 22)

Story (4) Long Told with no pauses

The Jones family went camping. On the way, their car ran out of gasoline. Instead of going where they planned, they camped in a small forest. They ended up having a BBQ and the children danced in front of the fire. It was a beautiful place, and it turned into a wonderful vacation.

(53 words; Total score 24)

Appendix B

Idea units in the four stories.

Story 1

last year	said	he did this	for a few weeks	only
John	he should	she taught him	he soon	high-calorie
started	make meals	some recipes	gave this up	dishes
university	on his own	he managed to	in the end	campus
his mother	to make sure	continue	he ate	in the cafeteria

___ / 25 Story 2				
there ~ be	in charge	no matter	the new	until
used to	always	the weather	cake	I left home
a cake shop	came up with	I went	every	I'll
in my town	new	there	weekend	remember it
the lady	cakes	to buy	I kept on doing this	as long as I live
___ / 26 Story 3				
British people	a sense of	pets	can tell	even
are fond of	well-being	don't take for granted	when their owners	sleep
keeping	taking care of	them	are sad	in the same bed
pets	them	dogs	it's no wonder	with their
they get	they say that	especially	some people	pets
___ / 25 Story 4				
Jones	ran out of	they camped	the children	place
family	gasoline	in a ~ forest	danced	turned into
went camping	instead of	small	in front of	wonderful
on the way	going	they ended up	the fire	vacation
their car	where they planned	having a BBQ	beautiful	
___ / 24				

References

Baddeley, Andrew D. 2015. Working memory in second language learning. In *Working Memory in Second Language Acquisition and Processing*. Edited by Zhisheng Wen, Mailce Borges Mota and Arthur McNeill. Bristol: Multilingual Matters, pp. 17–28.

Carrol, Gareth, and Kathy Conklin. 2020. Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech* 63: 95–122. [CrossRef] [PubMed]

Conklin, Kathy, and Norbert Schmitt. 2012. The Processing of Formulaic Language. *Annual Review of Applied Linguistics* 32: 45–61. [CrossRef]

Conklin, Kathy, and Rüdiger Thul. 2023. Word and multiword processing. In *The Routledge Handbook of Second Language Acquisition and Psycholinguistics*. Edited by Aline Godfroid and Holger Hopp. London: Routledge, pp. 203–15.

Deygers, Bart. 2020. Elicited imitation: A test for all learners? Examining the EI performance of learners with diverging educational backgrounds. *Studies in Second Language Acquisition* 42: 933–57. [CrossRef]

Gallagher, Karen, and Tamiko Azuma. 2018. Analysis of story recall in military veterans with and without mild traumatic brain injury: Preliminary results. *American Journal of Speech-Language Pathology* 27: 485–94. [CrossRef]

Haarman, Henk. 2013. Working Memory. 2012. In *Routledge Encyclopedia of Second Language Acquisition*. Edited by Peter Robinson. London: Routledge, pp. 694–99.

Hoang, Hien, and Frank Boers. 2016. Re-telling a story in a second language: How well do adult learners mine an input text for multiword expressions? *Studies in Second Language Learning and Teaching* 6: 513–35. [CrossRef]

Janssen, Niels, and Horacio A. Barber. 2012. Phrase frequency effects in language production. *PLoS ONE* 7: e33202. [CrossRef]
Kisser, Jason. E., Carrington R. Wendell, Robert J. Spencer, and Shari R. Waldstein. 2012. Neuropsychological performance of native versus non-native English speakers. *Archives of Clinical Neuropsychology* 27: 749–55. [CrossRef]
Kormos, Judit, and Anna Sáfár. 2008. Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition* 11: 261–71. [CrossRef]
Kostromitina, Maria, and Luke Plonsky. 2022. Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition* 44: 886–911. [CrossRef]
Lei, Yuyun, and Xun Yan. 2022. An Exploratory Study of Strategy Use on Elicited Imitation Tasks. *Frontiers in Psychology* 13: 917168. [CrossRef]
Lindstromberg, Seth, June Eyckmans, and Rachel Connabeer. 2016. A modified dictogloss for helping learners remember L2 academic English formulaic sequences for use in later writing. *English for Specific Purposes* 41: 12–21. [CrossRef]
Lundell, Fanny Forsberg, and Christina Lindqvist. 2013. Lexical aspects of very advanced L2 French. *The Canadian Modern Language Review* 70: 28–49. [CrossRef]
Martinez, Ron, and Norbert Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33: 299–320. [CrossRef]
McManus, Kevin, and Yingying Liu. 2020. Using elicited imitation to measure global oral proficiency in SLA research. A close replication study. *Language Teaching* 55: 116–35. [CrossRef]
Nation, I. S. Paul, and Azusa Yamamoto. 2012. Applying the four strands to language learning. *International Journal of Innovation in English Language Teaching and Research* 1: 167–81.
Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company.
Paquot, Magali, and Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistic* 32: 130–49. [CrossRef]
Peters, Elke. 2014. The effects of repetition and time of posttest administration on EFL learners' form recall of single words and collocations. *Language Teaching Research* 18: 75–94. [CrossRef]
Qian, David D., and Linda. H. F. Lin. 2020. The relationship between vocabulary knowledge and language proficiency. In *The Routledge Handbook of Vocabulary Studies*. Edited by S. Webb. London: Routledge, pp. 66–80.
Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In *Formulaic Sequences*. Edited by Norbert Schmitt. Amsterdam: John Benjamins, pp. 127–51.
Siyanova-Chanturia, Anna, and Niels Janssen. 2018. Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44: 2009–18. [CrossRef]
Solon, Megan, and Hae In Park. 2024. Elicited imitation in second language acquisition research: New insights to advance methodological rigor (Introduction to the special issue). *Research Methods in Applied Linguistics* 3: 1–4. [CrossRef]
Solon, Megan, Hae In Park, Carly Henderson, and Marzieh Dehghan-Chaleshtori. 2019. Revisiting the Spanish elicited imitation task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition* 41: 1027–53. [CrossRef]
Sonbul, Suhad. 2015. Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition* 18: 419–37. [CrossRef]
Stålhammar, Jacob, Per Hellström, Carl Eckerström, and Anders Wallin. 2022. Neuropsychological test performance among native and non-native Swedes: Second language effects. *Archives of Clinical Neuropsychology* 37: 826–38. [CrossRef] [PubMed]
Wolter, Brent. 2020. Key Issues in teaching multiword items. In *The Routledge Handbook of Vocabulary Studies*. Edited by Stuart Webb. London: Routledge, pp. 493–510. [CrossRef]
Yang, Pi-Lan. 2015. Interaction of working memory capacity in foreign language proficiency. *Concentric: Studies in Linguistics* 41: 95–115.
Yi, Wei, and Yanlu Zhong. 2023. The processing advantage of multiword sequences: A meta-analysis. *Studies in Second Language Acquisition* 46: 1–26. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Exploring Potential Factors Affecting Reading Comprehension in EAL Learners: A Preliminary Corpus-Based Analysis

Gavin Brooks [1,*], Jon Clenton [2] and Simon Fraser [3]

[1] Faculty of Foreign Studies, Kyoto Sangyo University, Kita-ku, Kamigamo, Motoyama, Kyoto 603-8555, Japan
[2] Graduate School of Humanities and Social Sciences, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Japan; jclenton@hiroshima-u.ac.jp
[3] Institute for Foreign Language Research and Education, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Japan; fraser@hiroshima-u.ac.jp
* Correspondence: gavinbrooks@gmail.com

Abstract: This article presents a study examining the vocabulary knowledge of English as an additional language (EAL) learners in two international schools in Japan in relation to the vocabulary profiles of the textbooks they are required to use in the classrooms. The vocabulary knowledge of 139 participants from two international schools was assessed using either the New Vocabulary Levels Test (NVLT) or the Updated Vocabulary Levels Test (UVLT). These results were compared to a 15 million-word corpus compiled from representative subject-specific textbooks to estimate the vocabulary coverage participants are likely to achieve. The findings revealed that EAL learners consistently scored lower than a combined group of their first-language English (FLE) and proficient L2 (PL2) peers, with fewer than 25% of EAL learners mastering the AWL before Grade 12. Furthermore, even the most frequent 5000-word bands provided only 91–93% coverage for subjects like biology and chemistry, leaving many EAL learners struggling to comprehend these texts. This analysis highlights the potential difficulties EAL learners may face in understanding the textbooks that are being used in EAL classrooms, underscoring the need for better vocabulary scaffolding and support for such learners in the international school context.

Keywords: English as an additional language; English-medium instruction; vocabulary; corpus; coverage

1. Introduction

The growth of English as an additional language (EAL) education worldwide (observed by, e.g., Smith, 2015) has led to an increase in EAL learners in countries such as the UK (Murphy & Unthiah, 2015) and the US (US Department of Education, National Center for Education Statistics, 2022), highlighting the need for effective EAL language instruction. Previous studies that have focused on the needs of EAL learners (Hessel & Murphy, 2019; Spencer et al., 2017; Townsend et al., 2016) have shown that English academic vocabulary is one vital area where teachers can support EAL learners. While there have been fewer studies conducted in the international school context, existing research indicates that EAL learners in this context also often lack the vocabulary skills to engage with the academic material in their classes (Coxhead & Boutorwick, 2018).

The issue of providing research-backed support for learners studying in an English as a medium of instruction (EMI) context in Japan is urgent, given the significant increase in the number of international schools operating in the country. Globally, International Baccalaureate (IB) schools experienced a 33.3% increase from 2016 to 2022, with Japan

seeing commensurate levels of growth during this period (International Baccalaureate, 2023). International schools in Japan often comprise learners from diverse L1 backgrounds with varying degrees of English language proficiency (Brooks et al., 2021), making it challenging for teachers to support EAL learners in the classroom. In international schools, the need for this support becomes even more crucial. This is because the challenges posed by learners are compounded by the fact that they speak multiple languages, and some are required to learn both English and the language of the wider community where the school is located (Carder, 2007). Consequently, learning English becomes significantly more complex in this context. The English language proficiency of EAL learners can be greatly affected by this, causing them to struggle to understand classroom materials in Japan (Brooks, 2023).

To better support the needs of EAL learners in the classroom, it is necessary to understand both the range of vocabulary with which EAL learners are likely to be familiar across different educational stages as well as the vocabulary composition of academic materials these students are expected to engage with during their studies. This study investigates the gaps in vocabulary knowledge that EAL learners studying in the International School context may demonstrate across grade levels and subjects. To explore EAL learners' vocabulary proficiency across various grade levels, we compared the vocabulary knowledge of two cohorts of learners enrolled in the international school context in Japan against the vocabulary profiles of the textbooks prescribed for their classes.

2. Literature Review

2.1. The Importance of Vocabulary for EAL Learners

While not a homogeneous group, EAL learners are typically defined as students who study in classrooms where English is the medium of instruction and who speak a language other than English at home (Murphy, 2014; Murphy & Unthiah, 2015). Sharples (2021) highlights two key implications from this definition: (i) the mainstream classroom is the primary place of learning for these students, and (ii) for them, language learning is inseparable from content learning. Additionally, research indicates that there are numerous factors which can hinder the academic success of EAL learners in EMI classrooms (Afitska & Heaton, 2019). This includes a considerable number of studies that suggest language difficulties play a critical role in their academic struggles (e.g., Clegg & Afitska, 2011; Coxhead & Boutorwick, 2018). In short, it is crucial for EAL learners to receive sufficient language support in the classroom. Without this support, they can face challenges in both their academic and linguistic development.

One area of language that research has consistently shown EAL learners struggle with is vocabulary knowledge (Faitaki et al., 2022). This is significant because research also shows that vocabulary is a key predictor of reading comprehension, not only for EAL learners (e.g., Brooks et al., 2021; Melby-Lervåg & Lervåg, 2014) but also for first-language English (FLE) learners (e.g., Ouellette & Beers, 2010; Tunmer & Chapman, 2012). Research shows (NALDIC, 2015) that EAL learners typically enter the educational system with lower levels of vocabulary knowledge than FLE learners studying at the same grade level. EAL learners have also been shown to take longer to learn and master the vocabulary they need for the classroom (Coxhead & Boutorwick, 2018) and, as a result, often have lower vocabulary levels than their FLE classmates across all grade levels (e.g., August et al., 2005; Murphy & Unthiah, 2015). Limited vocabulary knowledge can pose difficulties for EAL learners when it comes to understanding what they read, particularly regarding the textbooks they are assigned to read for their courses (Brooks et al., 2021; Marianne & Coxhead, 2023). Leung (2014) notes that textbooks can be particularly challenging for EAL learners because they often make use of precise vocabulary and specialised terminology and discuss unfamiliar

content. While the exact amount of vocabulary EAL learners require for academic success in an EMI setting is uncertain, studies indicate that two types of vocabulary knowledge are crucial for this group. These include both general academic vocabulary as well as high-frequency words (Coxhead & Boutorwick, 2018).

2.2. Supporting EAL Learners in the Classroom

To ensure the success of EAL learners in the classroom, it is crucial that they understand the texts they are required to read. Due to the multitude of skills learners need to master to comprehend texts in English, reading comprehension can be a difficult skill to assess (Melby-Lervåg & Lervåg, 2014). Factors that can impact a learner's ability to understand a text include prior linguistic knowledge (Droop & Verhoeven, 2003) as well as reading ability in their L1 (Chuang et al., 2012). Researchers have also found a significant relationship between learners' phonemic skills and their reading comprehension (Melby-Lervåg et al., 2012). While other factors are undoubtedly important for comprehending classroom textbooks, as discussed above, vocabulary knowledge has consistently been shown to be one of the key predictors of reading comprehension (Brooks et al., 2021). Consequently, it is important that teachers provide EAL learners with the vocabulary support necessary to comprehend the texts they are being assigned to read.

Multiple researchers have highlighted a correlation between vocabulary knowledge and reading comprehension (Brooks et al., 2021; Laufer, 1989; Schmitt et al., 2011). These same studies provide evidence that there is a threshold of coverage that is necessary for textual understanding (Laufer & Ravenhorst-Kalovski, 2010). Put simply, if learners cannot comprehend above a minimum threshold of the vocabulary in a text, they will struggle to grasp its overall meaning. It is frequently cited that comprehending approximately 95% of the vocabulary is necessary for learners to have an adequate understanding of what they are reading (Laufer, 1989). Students who are familiar with 95% of the words in a text may still need assistance with unfamiliar vocabulary while reading, such as asking a fellow student or teacher for help or referring to a dictionary when needed. The likelihood of learners being able to read a text independently increases if they are familiar with 98% of the words. Although having a wide range of vocabulary (for example, knowledge of sufficient vocabulary to provide 98% to 100% coverage of the words in a text) does not guarantee complete understanding, research consistently shows that an understanding of over 95% of the words in a text is a necessary, if not sufficient, condition for understanding (e.g., Kremmel et al., 2023; Schmitt et al., 2011). Of course, the question then becomes, what words do learners need to know to be able to understand a certain text? To effectively provide the type of vocabulary support that EAL learners need, it is essential to have a clearer understanding of the words they require to comprehend the reading material they are given.

2.3. Investigating Vocabulary in Learning Materials

Word lists are one important tool that teachers can use to help them identify the vocabulary that would be most pertinent to a specific group of learners or a given learning context. The most common type of word lists being used in the classroom are frequency-based lists such as university-level academic word lists (e.g., the Academic Word List (AWL; Coxhead, 2000) and the Academic Vocabulary List (AVL; Gardner & Davies, 2014)), middle school word lists (e.g., the Middle School Vocabulary List (MSVL; Greene & Coxhead, 2015)), and word lists for specialised subjects such as medicine and pharmacology (e.g., the Pharmacology Vocabulary List (PVL; Fraser, 2010) and the Medical Academic Vocabulary List (MAVL; Lei & Liu, 2016)). Frequency is a useful criterion on which to base a word list because it allows educators and researchers to judge how likely learners are to encounter a

given word, and it has been shown to be a good proxy for word difficulty (Benjamin, 2011; De Clercq & Hoste, 2016; Hancke et al., 2012). Studies support the use of frequency as a measure of word difficulty. For example, Brysbaert et al. (2011) found that word frequency was the most important factor for determining difficulty and, along with similarity to other words and word length, accounted for 40.5% of the variance in lexical decision times, whereas all of the other variables looked at in the study accounted for less than 2%.

A commonly employed technique for arranging frequency-based word lists is to divide the vocabulary into bins or groups based on frequency, usually utilising bins of 1000 words (Laufer & Nation, 1995). Examples of such an approach can be seen in popular word lists, such as the BNC/COCA (Nation, 2020) and the JACET 8000 list (Uemura & Ishikawa, 2004). Directional binning approaches work well for general word lists because they usually clearly indicate what the most important high-frequency words are for beginning learners (Flor et al., 2024). However, depending on the corpus being used, the coverage of such lists may not be as effective for mid- to low-frequency or domain-specific and technical words (Dang, 2019). These words are of particular importance to EAL learners as they often comprise the words learners need to know to be able to understand the topics they are learning about in their textbooks (Marianne & Coxhead, 2023). However, because there are no EAL-specific word lists, EAL teachers often need to rely on word lists compiled for FLE learners (e.g., Living Word Vocabulary, Dale & O'Rourke, 1976; EDL Core Vocabulary, Taylor et al., 1989), academic word lists designed for university learners (e.g., the AWL, Coxhead, 2000), or general word lists (e.g., the General Service List (GSL; West, 1953)) in the classroom. To determine how effective such lists will be at providing support in the EAL context, it is first necessary to better understand what coverage they provide over textbooks EAL learners are being asked to read, as well as how many new words learners need to acquire to achieve this coverage.

Researchers have used corpora to identify the types of words learners need to engage with domain-specific texts in various contexts (Coxhead, 2017). For example, Coxhead et al. (2016) asked students what words were necessary to study carpentry and compared their responses with technical terms extracted from a corpus of written carpentry texts, emphasising the importance of specialised vocabulary for academic success. Similarly, Miller and Biber (2015) outlined techniques for developing subject-specific word lists using a corpus of psychology textbooks. Additionally, Coxhead (2012) explored how corpora can aid educators and researchers in identifying the common words and phrases high school students are likely to encounter in their textbooks, demonstrating the value of corpus analysis in the educational context.

2.4. Vocabulary in Textbooks

One important area of research related to supporting learners in the classroom is analysing the vocabulary load of textbooks to evaluate whether the vocabulary aligns with learners' needs and to determine the comprehensibility of texts for students in different settings. While there are different ways of approaching how to analyse a textbook, one of the most common is to look at the lexical coverage of the text (Sun & Dang, 2020). Vocabulary load is linked to coverage, meaning the percentage of words learners are likely to know, which affects their ability to comprehend the material (Nation, 2005; Webb & Nation, 2008). Most of this research has focused on textbooks that international publishers have produced for use in English as a Foreign Language (EFL) classrooms (Benson & Madarbakus-Ring, 2021; Hsu, 2009; Matsuoka & Hirsh, 2010; O'Loughlin, 2012).

While the lexical loads of textbooks used in the university EFL classroom can be expected to vary across levels, more recent studies have provided support for the findings of Matsuoka and Hirsh (2010) that 95% coverage of EFL textbooks can be achieved

with around two to three thousand-word families. In the Japanese context, Benson and Madarbakus-Ring (2021) found that 95% coverage of a commercial textbook used at the university level in Japan could be achieved with the first 3000 words of the BNC/COCA. Yang and Coxhead (2022) looked at textbooks in use in China and found that 95% coverage was provided by the first 4000 words of the BNC/COCA. Hsu (2009) examined 36 different commercially available textbooks used in classrooms in Taiwan and found that learners would need between 2500 to 13,000-word families to achieve 95% coverage of the texts. This highlights the inconsistencies between the lexical demands. These findings are important, as they show that, while mastery of the first 3000 words would be sufficient for understanding most EFL university textbooks, this is not true in all cases. Teachers should take their learner's lexical proficiency into account when selecting a textbook.

High school textbooks in EFL contexts have been studied to determine their vocabulary load and how well they align with students' knowledge. Nguyen (2021) found that high school students in Vietnam required 3000–5000-word families to comprehend most of their textbooks. Sun and Dang (2020) analysed Chinese high school textbooks, showing students needed up to 9000-word families to reach 98% coverage. In both studies, the vocabulary knowledge of the learners was compared to the vocabulary loads of the textbooks; both found that most, if not all, of the learners would be unlikely to know enough words to achieve 95% coverage. These findings highlight a significant gap between learners' vocabulary knowledge and the lexical demands of their textbooks.

While these studies do provide us with some insight into the types of vocabulary learners need to know in an EFL setting, it is important to remember the way that textbooks are used in an EMI or CLIL classroom is different from how they are used in an EFL classroom (Sun & Dang, 2020). Where the focus of the lesson is the content and not the language, the vocabulary demands of the textbooks differ as EFL classes tend to focus on basic interpersonal communicative skills (BICS). In contrast, EMI classes focus on cognitive academic language proficiency (CALP, Cummins, 1979). In EFL contexts, the goal of the textbook is to expose learners to everyday language (BICS) and provide support for activities and discussions carried out in the classroom. The textbooks used in EMI and content-based classes, on the other hand, use academic and subject-specific language, and learners need to understand the content of the textbook to answer domain-specific questions on exams and quizzes. This is supported by studies that have shown the degree to which students studying in an EMI context outperform their EFL counterparts is higher for academic vocabulary than for everyday language due to their exposure to academic and subject-specific language in both their textbooks and the classroom (Castellano-Risco et al., 2020).

Outside of the EFL context, there have been a number of studies that have looked at the relationship between vocabulary and textbooks in EMI classrooms, both in English-speaking countries (e.g., Coxhead et al., 2015; Luxton et al., 2017) and in the international school settings (Coxhead & Boutorwick, 2018). There have also been a number of studies that have focused specifically on an analysis of the vocabulary found in textbooks written for content-based classes at both the junior high school (Greene & Coxhead, 2015) and high school levels (Coxhead & Boutorwick, 2018; Coxhead et al., 2010; Green & Lambert, 2018). Unlike the EFL studies discussed above, these studies focus on subject-specific textbooks written in English, such as those for science, mathematics, and literature, rather than textbooks that were written specifically for the purpose of learning English. Coxhead et al. (2010), for example, examined subject-specific textbooks (e.g., science textbooks in New Zealand secondary schools) and found that learners needed over 4000-word families for 95% coverage and over 14,000-word families for 98%. However, corpus-based studies that focused specifically on EAL learners and the unique vocabulary challenges they face remain under-explored in the literature. Expanding research in these areas would

provide a clearer understanding of the specific vocabulary demands and learning processes in EAL settings, ultimately contributing to more effective language instruction.

3. The Current Study

This study expands on prior research by examining the vocabulary knowledge of EAL learners in an international context in Japan and comparing it to the vocabulary found in the textbooks these learners use in class using the BNC/COCA word list (Nation, 2020). The aim is to better understand the vocabulary necessary for learners to comprehend these textbooks and assess the coverage provided by their current vocabulary knowledge. The specific goal is to gain insight into the vocabulary used in these textbooks to identify potential challenges EAL learners may face while reading them. As such, this study addresses four research questions:

1. To what extent does the vocabulary knowledge of English as an additional language (EAL) and first-language English (FLE) learners studying in an international school context vary across different grade levels?
2. To what extent does the vocabulary in subject-specific textbooks at an international school align with BNC/COCA frequency bands, and how does this alignment vary across International Baccalaureate (IB) subjects?
3. To what extent can EAL learners be expected to comprehend the vocabulary encountered in textbooks used for different IB subjects?
4. Using the BNC/COCA, how much additional vocabulary would EAL learners need to acquire to understand the language used in textbooks they would typically encounter within the international school context?

4. Methodology

4.1. Design

The study compared the vocabulary knowledge of EAL and FLE learners across seven different grade levels. A total of 139 participants were included in the final analysis, representing a diverse linguistic background. Each group's vocabulary knowledge was assessed using one of two versions of the Vocabulary Levels Test (VLT). The scores they received on the VLT were compared to an analysis of the vocabulary coverage in the textbooks used by the participants in their International Baccalaureate (IB) subjects.

The independent variables in the analysis were learner type (with two levels: (1) FLE and PL2 learners, and (2) EAL learners) and school level (with two levels: pre-IB, which covers Grades 6 to 9, and IB, covering Grades 10 to 12). The dependent variable was vocabulary knowledge, as measured by the VLT scores across the first five 1000-word bands. The analysis included an interaction term to examine whether the difference in vocabulary knowledge between FLE/PL2 and EAL learners varied across school levels.

The corpus analysis focused on five subject areas (literature, maths, physics, chemistry, and biology), examining the frequency and coverage of vocabulary in textbooks used in these subjects. Vocabulary comprehension was measured in terms of how well the learners' existing knowledge aligned with the vocabulary demands of these textbooks.

4.2. Participants

The study took place at two different International Schools in Japan. A total of 142 participants (N = 142) from diverse linguistic and cultural backgrounds participated in the study. While Japanese and English were the two most common languages spoken, 22 different languages were represented in our dataset. We used a Rasch analysis (Beglar, 2010) to evaluate participant scores on the uVLT and the NVLT. The majority of VLT scores demonstrated a good fit for the Rasch model. However, we removed three learners with

very high outfit scores (*Zstd* > 8.76) from the dataset. After the three participants were excluded from the final analysis, the final dataset consisted of 139 participants (68 male and 71 female learners ranging in age from 13 to 18 years old).

Following the procedure used in previous studies (Coxhead & Boutorwick, 2018), we categorised the participants as either FLE, Proficient L2 (PL2) learners, or EAL learners based on nationality, time spent in English-speaking countries, languages spoken at home, and teacher assessments of English proficiency. For this study, EAL learners were defined as those who did not speak English at home and required additional language support in the classroom. FLE learners were participants who were both proficient in English and either spoke English predominantly at home or had significant exposure to English-speaking environments. PL2 learners were proficient in English but primarily spoke a language other than English at home. The term PL2 was chosen over non-native speaker (NNS) to avoid the deficit model associated with native and non-native speaker labels. Table 1 details how many EAL, PL2, and FLE participants there were at each grade level.

Table 1. Breakdown of FLE, PL2, and EAL participants by grade.

Grade	6th	7th	8th	9th	10th	11th	12th	Total
FLE	0	3	2	1	1	3	1	11
PL2	2	2	4	6	3	6	2	25
EAL	4	13	15	17	20	28	6	103
Total	6	18	21	24	24	37	9	139

Given that our study was primarily concerned with the vocabulary sizes of EAL learners and investigating the relationship between their vocabulary knowledge and their ability to engage with and comprehend classroom readings, as highlighted in previous studies (Coxhead & Boutorwick, 2018; Marianne & Coxhead, 2023; Murphy & Unthiah, 2015), we made the decision to group the FLE and PL2 learners for the purpose of analysis. This was based on several factors. Firstly, this approach mirrored that of earlier studies (Brooks, 2023; Coxhead & Boutorwick, 2018) that have focused on EAL learners and compared their vocabulary knowledge and progress with those of their FLE and PL2 peers. Secondly, teachers considered both groups to be similar in the classroom context, and additional language support, such as pull-out classes, was reserved exclusively for EAL learners. Thirdly, demarcating between FLE and PL2 was challenging as the responses they gave to the questions on the surveys were very similar. For example, a Grade 10 student who had lived in the US for the first eight years of his life but spoke Japanese at home selected Japanese as his first language, while another learner who had bilingual parents but spoke predominantly English at home and had never lived overseas identified English as her L1. Despite initially being in different groups, the learners from the FLE and PL2 groups shared the most similarities in both their responses to the survey questions as well as how their language skills in the classroom were evaluated by their teachers. Given all these factors, it made sense to combine the FLE and PL2 learners, allowing us to focus on comparing the EAL learners with the rest of the cohort more effectively.

4.3. Materials

4.3.1. The Vocabulary Levels Tests

The learners' vocabulary knowledge was measured using the Vocabulary Levels Test (VLT). The VLT is a standardised assessment instrument designed to evaluate a learner's vocabulary knowledge across different frequency bands to gain a deeper understanding of their language proficiency (Schmitt et al., 2001). In this study, the participants were

given one of two different versions of the VLT test. Both tests cover the first five 1000-word bands of the BNC/COCA. These tests were chosen over previous VLTs (e.g., Nation, 1983; Schmitt et al., 2001) because they cover a greater number of frequency bands and are based on a more recently compiled set of word lists, the BNC/COCA (Nation, 2020). In total, 49 participants (31 EAL learners, 6 FLE learners, and 12 PL2 learners) took the new Vocabulary Levels Test (NVLT) and 90 participants (72 EAL learners, 5 FLE learners, and 13 PL2 learners) took the updated Vocabulary Level Test (uVLT).

We administered the new Vocabulary Levels Test (McLean & Kramer, 2016) to the first group. This vocabulary assessment tool is a multiple-choice test consisting of 24 questions for each of the first five bands. The test also includes an academic vocabulary section based on Coxhead's (2000) AWL. Each of the individual items consists of the target word, which is given by itself and in context, along with four possible responses. The examinee has to select the correct response, which can be either a single word or a phrase closest in meaning to the target word (see Figure 1).

1. time: They have a lot of **time**.
a. money
b. food
c. hours
d. friends

Figure 1. A sample question from the NVLT. Adapted from McLean and Kramer (2016), page 4.

We administered Webb et al.'s (2017) updated Vocabulary Levels Test to the second group of participants. As with the NVLT, this test measures knowledge of the first five 1000-word bands. The uVLT also requires examinees to match a word with its definition or explanation. However, it differs from the NVLT in that the items are not presented individually but grouped into clusters, with each cluster containing three related words. There are 10 clusters for each of the first five 1000-word bands. The test taker must match each word to its corresponding definition or explanation (see Figure 2). The uVLT includes a representative portion of nouns, verbs, and adjectives (15, 9, and 6 items per level, respectively) selected from each of Nation's (2020) BNC/COCA bands. Because the uVLT does not include a section for Coxhead's (2000) Academic Word List, we supplemented the test with the AWL section of the NVLT.

	eye	father	year	van	voice	night
body part that sees	✓					
parent who is a man		✓				
part of the day with no sun						✓

Figure 2. A sample question from uVLT, along with the answers provided. Adapted from Webb et al. (2017), page 61.

We understand that it would have been better to use the same assessment tools for both groups of learners. However, the assessments were conducted as part of two larger studies (e.g., Brooks et al., 2021) and, in the context of this paper, we felt that it was important to include the data from both assessments as that allows us to examine larger trends within the population. The tests serve as a tool for assessing the learner's comprehension of the frequency band, and the effectiveness of both assessments in demonstrating the mastery of frequency bands by learners has been proven in studies conducted by the authors of each assessment tool (McLean & Kramer, 2016; Webb et al., 2017) and have been used in this capacity by other researchers (e.g., Ha, 2021; Kremmel et al., 2023; Xodabande &

Hashemi, 2023). It is crucial to clarify that these scores are not being compared to any other scores or utilised as a measure of linguistic proficiency but are being used to provide us with a picture of the level of vocabulary knowledge EAL learners in this context are likely to possess. Given this, we feel that any potential variance between the tests would not invalidate the picture they are able to give us of the average level of vocabulary knowledge of the different groups of students in this study. Although there may be a slight variation between the tests, the benefits of gathering information about the vocabulary knowledge of a larger number of students in diverse settings outweigh any potential drawbacks related to differences between the assessments themselves.

4.3.2. The Corpus

To create a corpus representative of the textbooks used by learners in the classroom, we compiled a set of domain-specific corpora sourced from a variety of textbooks taken from across the different subjects the participants were studying at the two schools. The corpus comprises five subjects (Table 2): literature, maths, physics, chemistry, and biology. The corpus focuses on the textbooks used during the IB diploma and does not include the handouts and textbooks the participants were using at the middle school level. The rationale for this was two-fold. Initially, these textbooks aided in building a strong vocabulary foundation for EAL learners in the IB program. Secondly, teachers at the middle school level used handouts and teacher-created materials more frequently in the classroom rather than relying on prescribed textbooks. While there were no middle-school textbooks in the corpus, we feel that the vocabulary in the corpus is relevant to these grades because, in the IB setting, one of the primary goals of the middle-school grades is to prepare students for the IB program. Part of this preparation involves having students learn the vocabulary that they need to succeed in the IB context.

Table 2. Total number of texts and tokens for each subject.

Subjects	Total Textbooks	Total Tokens
Literature	7	1,122,212
Maths	8	942,784
Physics	5	1,135,845
Chemistry	5	1,315,730
Biology	5	941,215
Total	30	5,457,786

We digitised the textbooks by scanning them into the computer, then meticulously reviewed and refined the scanned texts to rectify any errors that might have occurred during the scanning process. This thorough process was conducted using a text editor, following the procedures suggested by Nation (2016). We utilised R and Excel to deal with errors not addressed during the initial cleaning phase. All of the errors that resulted in non-words were cleaned by exporting a csv of all the off-list words and manually checking those against the original PDFs. This was carried out by the primary researcher and a group of research assistants. However, despite the extensive data cleaning, it was not possible to completely eliminate all noise from such a large corpus, and some OCR (optical character recognition) errors remain, such as those with mathematical and chemical formulas, with numbers such as “1” being recognised as an “I”, or with punctuation. Although this may restrict the use of data to analyse features of the text such as paragraph length or mean sentence length, the focus on removing incorrectly scanned words means that the impact of this noise on vocabulary measures was minimal.

4.4. Procedures

As noted above, the participants for the study were recruited from two international schools in Japan. This was carried out through a collaboration with the EAL teachers and the principals at both schools, which helped to facilitate the recruitment process. All students in grades 6 through 11 were administered the Vocabulary Levels Tests (VLTs) as part of their regular classroom activities. Due to time constraints, not all grade 12 students were able to participate, as they were heavily engaged in preparing for their International Baccalaureate (IB) exams and final assessments.

At the start of the project, we obtained consent from both the participants and their parents or guardians. This process was carried out collaboratively with the research team, the students, the parents, and the school administration. Consent was required before the administration of the VLT, and only students from whom consent was received were included in the final dataset.

The VLT was administered in the classroom by the students' regular teachers. The test was paper-based, and students were given one hour to complete the VLT along with a short survey on their language background. All participants finished within the allotted time.

At the same time they took the VLT, participants also filled in a short survey which included questions on nationality, languages spoken at home, and time spent overseas. Additionally, students were asked to self-identify their first language (L1), with the option to select multiple L1s, acknowledging the number of bilingual and multilingual learners in the study.

5. Analysis

5.1. Analysing the VLTs

Following previous studies (Coxhead & Boutorwick, 2018), participants who scored above 86% on the words from a single frequency band were considered to have mastered that frequency band. The number of participants at each grade level that achieved mastery was used to determine how likely participants were to be able to read the required textbooks for each of the IB subjects. The overall VLT scores were used for the between-group comparisons at each grade level.

5.2. Analysing the Corpus

The texts that made up the corpus were initially imported and preprocessed in R (v4.2.1; R Core Team, 2021). After this, we lemmatised and identified proper nouns in the corpus using Python in conjunction with spaCy, a widely used natural language processing library that allows for efficient and accurate linguistic annotations, including tokenisation, part-of-speech tagging, lemmatisation, and named entity recognition. Following these processing steps, the annotated data were exported as a CSV file. The final analysis of the lemmatised corpus was conducted in R, utilising the tidyverse package.

Once we had a cleaned and lemmatised version of the corpus, we used a pre-prepared data frame created from Nation's (2020) BNC/COCA word family lists (from 1000 to 25,000) to determine the frequencies of word families in the text. During this stage, we also made use of four supplementary lists from Nation (2020) to identify proper nouns (e.g., Europe, David), abbreviations (e.g., pm, rpm, kwh), compounds (e.g., birthday, airbag), and marginal words (e.g., ah, mm, X, Y). These lists were supplemented with the list of proper nouns that spaCy had identified in the corpus during the previous stage.

In the next phase of our research, we further analysed the off-list words to ensure that all abbreviations, marginal words, and proper nouns from the text were accurately added to the appropriate supplementary list. To achieve this, we checked each of the off-list

words against the original PDF to verify that the OCR had been carried out correctly on that word and that there were no errors in the final text files. Any errors that we identified were manually corrected. Words that were correct and were either abbreviations, marginal words, or proper nouns were added to the appropriate list. After these amendments, we used the updated texts and lists to rerun our analysis. We then used this final analysis to determine what coverage the different frequency bands of the BNC/COCA could provide over the five subcorpora.

6. Results

Research Question 1: To what extent does the vocabulary knowledge of English as an additional language (EAL) and first-language English (FLE) learners in an international school context vary across different grade levels?

An examination of the VLT scores (Table 3) shows that the majority of the participants were not able to attain mastery of the AWL or the BNC/COCA mid-frequency bands. These findings reflect previous studies' findings (e.g., Coxhead & Boutorwick, 2018). Additionally, we noticed that a considerable portion of the participants struggled with learning the common high-frequency words from the 2000 range before reaching Grade 9, and less than half of both groups were able to grasp the 3000-word levels before Grade 12. The mastery of the AWL also posed a significant challenge for learners at all grade levels, with fewer than 25% of the participants being able to master these words until the 12th grade. Because of how important the general academic words found on the AWL are for comprehending academic texts such as school textbooks (Greene & Coxhead, 2015; Hu et al., 2021), the lower levels of proficiency the participants displayed on vocabulary items from the AWL would suggest that they would be likely to encounter difficulties reading texts appropriate for their grade level.

Table 3. Vocabulary mastery at each grade level (All participants).

Grade	1000	2000	3000	4000	5000	AWL
Grade 6 (n = 6)	100.00%	33.33%	16.67%	16.67%	16.67%	16.67%
Grade 7 (n = 18)	88.89%	50.00%	5.56%	27.78%	27.78%	5.56%
Grade 8 (n = 21)	95.24%	66.67%	33.33%	42.86%	38.10%	19.05%
Grade 9 (n = 24)	95.83%	83.33%	50.00%	45.83%	45.83%	20.83%
Grade 10 (n = 24)	87.50%	79.17%	25.00%	50.00%	45.83%	12.50%
Grade 11 (n = 37)	97.30%	86.49%	37.84%	48.65%	40.54%	27.03%
Grade 12 (n = 9)	100.00%	88.89%	55.56%	66.67%	44.44%	55.56%
Average (n = 139)	94.24%	74.82%	33.09%	44.60%	39.57%	20.86%

Note: AWL = Academic Word List.

We conducted separate analyses for EAL learners and PL2/FLE learners, highlighting that EAL participants showed significantly lower levels of proficiency with both high-frequency and mid-frequency vocabulary (Table 4) compared to PL2/FLE learners (Table 5). We found that the majority of PL2/FLE participants were able to master the first 5000 BNC/COCA frequency bands. In contrast, fewer than 50% of EAL participants could master beyond the 2000-word band. Even at grade 12, only 50% of EAL participants had mastered the 3000-word band, and just 33% had mastered the 5000-word band. However, despite their higher levels of vocabulary proficiency, we found that many of the PL2/FLE still struggled with the AWL, suggesting that vocabulary, particularly academic vocabulary, remains a challenge for even this more adept group of students.

Table 4. Vocabulary mastery at each grade level (EAL Learners).

Grade	1000	2000	3000	4000	5000	AWL
Grade 6 (n = 4)	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Grade 7 (n = 13)	84.62%	30.77%	0.00%	7.69%	0.00%	0.00%
Grade 8 (n = 15)	93.33%	53.33%	20.00%	26.67%	20.00%	0.00%
Grade 9 (n = 17)	94.12%	76.47%	29.41%	29.41%	23.53%	0.00%
Grade 10 (n = 20)	85.00%	75.00%	15.00%	40.00%	35.00%	10.00%
Grade 11 (n = 28)	96.43%	82.14%	17.86%	35.71%	25.00%	14.29%
Grade 12 (n = 6)	100.00%	83.33%	50.00%	50.00%	33.33%	50.00%
Average (n = 103)	92.23%	66.02%	18.45%	30.10%	22.33%	8.74%

Note: AWL = Academic Word List.

Table 5. Vocabulary mastery at each grade level (PL2/FLE learners).

Grade	1000	2000	3000	4000	5000	AWL
Grade 6 (n = 2)	100.00%	100.00%	50.00%	50.00%	50.00%	50.00%
Grade 7 (n = 5)	100.00%	100.00%	20.00%	80.00%	100.00%	20.00%
Grade 8 (n = 6)	100.00%	100.00%	66.67%	83.33%	83.33%	66.67%
Grade 9 (n = 7)	100.00%	100.00%	100.00%	85.71%	100.00%	71.43%
Grade 10 (n = 4)	100.00%	100.00%	75.00%	100.00%	100.00%	25.00%
Grade 11 (n = 9)	100.00%	100.00%	100.00%	88.89%	88.89%	66.67%
Grade 12 (n = 3)	100.00%	100.00%	66.67%	100.00%	66.67%	66.67%
Average (n = 36)	100.00%	100.00%	75.00%	86.11%	88.89%	55.56%

Note: AWL = Academic Word List.

The descriptive statistics of the scores for each of the Grade levels for the EAL and FLE/PL2 groups are given in Table 6.

Table 6. Descriptive statistics of the VLT scores for the EAL and FLE/PL2 across grade levels and frequency bands.

Grade and Group	n	VLT Score (Mean)	VLT Score (SD)	K1 (Mean)	K1 (SD)	K2 (Mean)	K2 (SD)	K3 (Mean)	K3 (SD)	K4 (Mean)	K4 (SD)	K5 (Mean)	K5 (SD)	AWL (Mean)	AWL (SD)
Grade 6 EAL	4	63.73	9.24	95.83	3.39	71.88	12.88	54.17	11.79	63.52	12.91	56.25	12.50	40.83	13.73
Grade 6 FLE/PL2	2	86.90	12.30	97.90	2.97	93.75	8.84	79.20	17.68	85.40	14.71	83.30	17.68	81.65	11.81
Grade 7 EAL	13	58.55	16.31	93.14	10.90	77.75	16.42	42.85	20.41	49.33	24.24	43.85	19.16	44.50	19.17
Grade 7 FLE/PL2	5	87.74	5.62	98.32	2.30	99.40	1.34	76.06	12.27	88.64	8.34	90.40	3.57	74.00	11.89
Grade 8 EAL	15	71.36	17.28	95.90	7.84	83.45	17.00	65.31	22.99	70.67	18.70	56.89	30.46	55.99	20.75
Grade 8 FLE/PL2	6	92.57	9.12	99.30	1.71	99.50	1.22	88.77	9.73	94.30	12.02	90.75	15.38	82.67	19.58
Grade 9 EAL	17	73.96	14.80	96.42	6.34	91.51	11.50	65.88	20.58	67.08	22.44	66.52	20.60	56.77	16.78
Grade 9 FLE/PL2	7	94.76	2.31	98.20	2.24	98.57	3.78	91.54	4.96	96.61	6.42	92.59	2.19	90.86	4.70
Grade 10 EAL	20	74.97	17.75	95.78	8.49	85.97	18.51	66.47	19.91	71.92	21.78	65.21	25.82	64.44	21.44
Grade 10 FLE/PL2	4	93.28	1.97	100.00	0.00	100.00	0.00	87.00	2.92	97.50	3.32	91.58	3.60	83.15	8.29
Grade 11 EAL	28	75.55	14.27	97.78	4.76	90.42	12.01	66.31	18.08	73.50	20.05	64.84	22.79	60.62	22.21
Grade 11 FLE/PL2	9	95.03	3.31	99.53	1.40	100.00	0.00	93.36	3.59	96.67	6.10	92.14	4.51	88.70	8.52
Grade 12 EAL	6	79.63	16.38	98.50	1.64	91.67	10.78	77.33	16.69	74.17	23.75	66.67	23.87	69.83	27.37
Grade 12 FLE/PL2	3	95.83	5.31	100.00	0.00	100.00	0.00	94.43	9.64	98.60	2.42	89.73	9.34	92.33	10.79

Note: K01 to K05 refer to the 1000 to 5000 word frequency bands; AWL refers to the Academic Word List.

A linear regression model was used to compare the vocabulary knowledge of EAL learners with that of FLE and PL2 learners across different grade levels (Table 7). Given the small number of participants at some grade levels, the participants were divided up into pre-IB, covering Grades 6 to 9, and IB, covering Grades 10 to 12. The results showed

that FLE/PL2 learners consistently outperformed EAL learners on the Vocabulary Levels Test (VLT). On average, FLE/PL2 learners scored about 19 percentage points higher than EAL learners. This significant difference in vocabulary knowledge was evident across both the pre-IB (Grades 6–9) and IB (Grades 10–12) levels. Additionally, EAL learners in pre-IB scored lower than those in the IB level, indicating an improvement in vocabulary knowledge as they advanced in grade levels. However, the gap between EAL and FLE/PL2 learners remained consistent across grade levels, suggesting that while all learners develop their vocabulary over time, EAL learners continue to lag behind their FLE/PL2 peers throughout their schooling.

Table 7. A summary of the linear regression model comparing vocabulary scores of FLE/PL2 and EAL learners across school levels.

	Estimate	Std. Error	t Value	Pr(>|t|)
Intercept	0.76	0.02	39.26	$<2 \times 10^{-16}$ ***
FLE/PL2	0.19	0.04	4.70	6.45×10^{-6} ***
Pre-IB	−0.08	0.03	−2.70	0.01 **
FLE/PL2 × Pre-IB	0.04	0.06	0.79	0.43

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Research Question 2: To what extent does the vocabulary in subject-specific textbooks at an international school align with BNC/COCA frequency bands, and how does this alignment vary across IB courses?

The vocabulary profiles of the domain-specific corpora (Table 8) show that the literature textbooks were likely to be the most accessible for EAL learners. Based on the coverage of the BNC/COCA over this corpus, we would expect learners to achieve the 95% coverage threshold required for comprehension if they could master the first five 1000-word bands from the BNC/COCA. The next most accessible group of textbooks were those from the maths corpus, where the coverage provided by the first five 1000-word bands was 94.28%, very close to the minimum threshold of 95%. However, for all the other corpora, the coverage provided by the first 5000 words fell significantly below the 95% mark.

Table 8. Coverage provided by the AWL and BNC/COCA by subject.

Frequency Band	Literature	Math	Physics	Biology	Chemistry	Average
2000	86.84%	83.79%	84.03%	77.08%	77.79%	81.91%
5000	96.55%	94.28%	91.48%	91.45%	92.60%	93.27%
AWL	6.87%	8.30%	7.36%	7.70%	8.60%	7.77%

Note: AWL = Academic Word List.

The first 2000 frequency bands provided notably low coverage, particularly in the chemistry and biology texts, where less than 78% of all tokens were from the first two bands. This is significantly lower than the coverage these bands have been found to provide over texts written specifically for English language learners, where the first 2000 words have been found to provide over 92% coverage (Sun & Dang, 2020). These findings suggest that EAL learners would struggle with these textbooks given their vocabulary mastery level.

Research Question 3: To what extent can EAL learners be expected to comprehend the vocabulary encountered in the textbooks used for different IB subjects?

We estimated the percentage of vocabulary in the textbooks that EAL learners would likely know by analysing their scores on the NVLT or uVLT. We then looked at the coverage provided by each of the bands of the BNC/COCA over the various domain-specific corpora

(Table 9). By looking at the coverage provided by the words from the frequency levels that an individual participant had displayed mastery of, we were able to determine the overall vocabulary coverage a learner would likely have over the texts from different subjects. Using these numbers, we were able to infer what percentage of words the participants would likely know from the different corpora.

Table 9. Expected coverage per discipline from the domain-specific subcorpora.

Category	Literature	Biology	Chemistry	Physics	Math
1000 word band	74.54%	65.51%	63.39%	70.88%	72.38%
2000 word band	9.56%	10.89%	12.94%	12.37%	9.96%
3000 word band	7.35%	9.23%	9.26%	7.58%	8.17%
4000 word band	1.52%	3.03%	3.78%	3.28%	2.71%
5000 word band	0.84%	2.15%	1.76%	1.28%	1.61%
6 to 25,000 word bands	2.33%	7.08%	5.96%	3.30%	3.05%
Other	3.86%	2.12%	2.90%	1.30%	2.12%

To assess whether the differences in coverage that learners were likely to achieve across subjects, as identified in the descriptive analysis, were statistically significant, a repeated-measures ANOVA was conducted, followed by pairwise comparisons using the Tukey adjustment to control for multiple comparisons. The ANOVA revealed a significant difference in coverage between academic subject, $F(4, 408) = 4858.89$, $p < 0.001$, $\eta^2 = 0.09$, showing that coverage learners would be likely to have over the different sub-corpora would vary between subjects. The subsequent post-hoc analysis showed that the largest difference was between biology and literature, with students achieving significantly higher coverage in literature ($M = 88.98$, $SE = 0.79$) than in biology ($M = 82.39$, $SE = 0.83$; Mean Difference = 6.58, $p < 0.001$). The only non-significant difference was between math ($M = 87.54$, $SE = 0.83$) and physics ($M = 87.54$, $SE = 0.85$; Mean Difference = 0.01, $p = 0.912$), suggesting learners would have similar vocabulary coverage across both of these two subjects. These findings suggest that, with one exception, the coverage provided by the BNC/COCA corpus is domain-specific, indicating that learners require varying levels of vocabulary support depending on the subject they are studying.

Figure 3 presents a histogram of the vocabulary coverage participants would likely know for each corpus. The data reveal that most EAL participants lack the vocabulary needed to read discipline-specific textbooks, with biology and chemistry being the most challenging subjects and maths and literature the easiest. Given the vocabulary levels displayed by EAL learners, even at the Grades 11 and 12 level, we expect them to struggle to fully comprehend the textbooks used in their classes.

Research Question 4: Using the BNC/COCA, how much additional vocabulary would EAL learners need to acquire to understand the language used in textbooks they would typically encounter within the international school context?

To answer this fourth and final question, we need to look at the gap that exists between the learners' current knowledge and the number of words they would need to learn, using existing word lists, to achieve 95% or greater coverage. From what we have discussed above, it is evident that learners often lack proficiency in frequency bands exceeding the first 2000 words. We also know that they are not likely to have mastered the Academic Word List. Since learners would have to master the 5000 to 9000 most frequently used words in the BNC/COCA to adequately cover various subjects, the number of words they would need to learn solely based on the BNC/COCA is likely too large for them to acquire within their available time (Table 10). For example, a typical 10th-grade EAL student would

need to learn around 7000 new words to achieve the threshold necessary to comprehend a biology or chemistry textbook successfully.

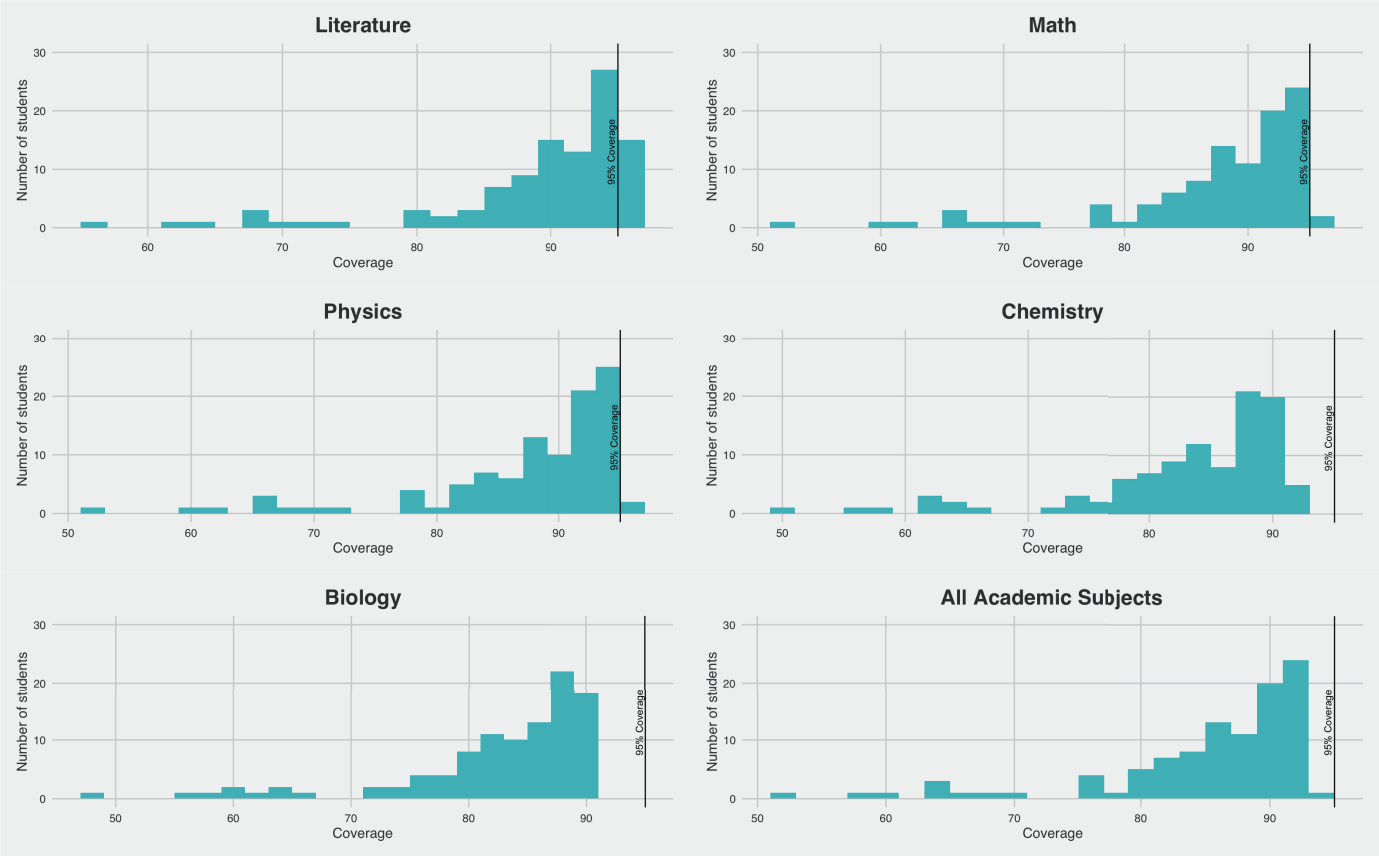

Figure 3. A histogram depicting the percentage of words EAL participants would likely understand in each of the subject-specific corpora.

Table 10. Mastery of the word frequency bands for individual participants.

Participant	1000	2000	3000	4000	5000	AWL
S04	**100%**	**97%**	80%	**87%**	**87%**	67%
S09	**100%**	**100%**	67%	77%	**90%**	63%
S12	**100%**	**97%**	70%	**87%**	83%	77%
S28	**100%**	**100%**	70%	**93%**	**87%**	73%
S32	**100%**	**97%**	67%	**100%**	73%	67%
S51	**100%**	**100%**	**87%**	83%	**87%**	77%

Note: AWL = Academic Word List. Bold text indicates participants who have shown mastery of that frequency band.

Examples from our study further illustrate this issue: participant S04 was able to master the 4000- and 5000-word bands but not the 3000-word band. Participant S09 mastered the 5000-word band but not the 3000- or 4000-word bands. According to previous studies (e.g., Schmitt & Schmitt, 2014; Teng, 2019), learners are likely to acquire words based on the number of exposures to those words. Given this fact, our finding suggests that, beyond the first 2000 high-frequency words, the BNC/COCA frequency list may not accurately represent the frequency of the vocabulary items in the textbooks that EAL learners are being asked to read for their classes. This misalignment calls into question the effectiveness of using traditional frequency lists for supporting EAL learners' vocabulary acquisition. Given this fact and the strong and significant correlation between vocabulary

knowledge and reading comprehension shown by other studies (e.g., Brooks et al., 2021; Laufer & Ravenhorst-Kalovski, 2010), more research is needed regarding EAL learners' vocabulary knowledge and needs in an EMI context. Additionally, there is a pressing need to develop tools that effectively meet these vocabulary needs.

7. Discussion

The current study was designed to examine the vocabulary knowledge of EAL learners in an international context in Japan and compare it to the vocabulary found in the textbooks these learners use in class, using the BNC/COCA word list. The aim was to gain insight into the vocabulary used in these textbooks to identify potential challenges EAL learners may face while reading them. The study was driven by four research questions, each of which we discuss in turn below.

Regarding Research Question 1, the data strongly support the findings of previous research (Coxhead & Boutorwick, 2018), indicating a widespread challenge among learners, especially those with English as an additional language, in mastering academic vocabulary. Similar challenges were noted in studies by Dixon et al. (2020), which found that while EAL learners do acquire vocabulary, they are often not able to do so with sufficient speed to keep up with their FLE-speaking peers. The finding that fewer than 25% of participants could master the AWL before Grade 12 is concerning, as academic vocabulary is essential for reading comprehension and success in the classroom (Brooks, 2023; Green & Lambert, 2018; Greene & Coxhead, 2015). Studies conducted by Coxhead and Boutorwick (2018) and Marianne and Coxhead (2023) yielded similar findings, indicating low levels of academic vocabulary mastery among EAL learners in comparable educational contexts. Given their difficulties with academic texts, it is probable that EAL learners will face substantial challenges in their academic pursuits, particularly as they progress to higher grades where the academic demands intensify.

For Research Question 2, the results reinforce the need for domain-specific vocabulary lists, a need that has been extensively discussed in earlier research (Green & Lambert, 2018; Greene & Coxhead, 2015). In line with previous studies on vocabulary (Coxhead, 2017), it appears that the difficulty of the vocabulary found in a corpus can vary markedly across domains. While more proficient learners may possess sufficient vocabulary resources to handle content in subjects like literature and mathematics, which were found to be relatively accessible with approximately 95% coverage achieved using the first 5000 BNC/COCA word bands, specialised subjects such as biology and chemistry present greater lexical challenges, resulting in a notable drop in coverage. This builds upon previous studies that have shown both the difficulty of the vocabulary in textbooks from a specific domain, such as the sciences (Hu et al., 2021), as well as the importance of this vocabulary because of the strong connection between content knowledge and technical vocabulary (Coxhead, 2017; Woodward-Kron, 2008). The results of our research corroborate the findings of previous studies and underscore the distinct obstacles posed by subject-specific terminology, emphasising the necessity of providing tailored lexical assistance to students grappling with such domains.

Although the 5000 most frequent words offer a wide range of coverage, the BNC/COCA corpus falls short in adequately capturing the specialised terminology prevalent in the fields of biology, chemistry, and physics. This deficiency echoes similar findings from prior studies on the vocabulary demands of English as an additional language (EAL) learners, highlighting the need for a more comprehensive approach to vocabulary acquisition in these subjects (Coxhead & Boutorwick, 2018; Coxhead et al., 2010). The low coverage in these specialised domains emphasises the difficulties faced by EAL learners, especially considering that only a minority (e.g., 33% of Grade 12 participants) demonstrated mastery

of the requisite vocabulary bands. This further underscores the importance of developing targeted vocabulary lists to support comprehension in more specialised academic texts, as earlier studies have suggested (Green & Lambert, 2018; Greene & Coxhead, 2015).

Research Question 3 investigates the vocabulary coverage of discipline-specific textbooks, analysing the expected coverage based on English as an additional language (EAL) learners' Vocabulary Levels Test (VLT) scores. This analysis reveals a concerning trend: most EAL learners possess insufficient vocabulary to fully comprehend the content within these textbooks. This supports earlier research, which has shown that these learners tend to struggle with reading comprehension in an academic setting (Brooks, 2023; Coxhead et al., 2010; Murphy & Unthiah, 2015).

While the EAL learners in this study demonstrated improvement in vocabulary knowledge across grade levels, the rate of this growth did not match the rapid gains observed in other researchers, such as Coxhead et al. (2015), who found that 15–16-year-old FLE speakers enrolled in New Zealand high schools learned an average of over 1300-word families per year. While our study indicated vocabulary knowledge improvements across grade levels, these improvements were less substantial than those observed in a study by Coxhead and Boutorwick (2018), where some EAL learners achieved comparable vocabulary knowledge levels to their native-speaking peers within a shorter timeframe—four years for high-frequency vocabulary and five years for academic vocabulary. Our research reveals a persistent gap in vocabulary knowledge, even among upper-grade learners, particularly in subjects like biology and chemistry. Even Grade 12 English as an additional language (EAL) learners show a significant lack of proficiency in essential vocabulary bands, highlighting a substantial discrepancy between their vocabulary comprehension and the demands of these disciplines. This presents a significant barrier to their academic success in these subjects (Marianne & Coxhead, 2023; Woodward-Kron, 2008).

Although the BNC/COCA list could potentially help address these vocabulary gaps, our analysis for Research Question 4 reveals that EAL learners' vocabulary profiles deviate significantly from the typical profile predicted by this frequency list (Nation, 2016), and further suggests that the number of words they would need to learn based on these lists is too extensive to be realistically managed. These lists organise words by how often learners encounter them in real-world contexts, with higher-frequency words appearing more frequently in texts. As a result, learners are expected to learn higher-frequency words sooner due to repeated exposure. However, previous studies have shown that textbooks in subjects like maths and science can have very different vocabulary profiles and loads than other texts, such as novels or English language textbooks (Groves, 2016; Hu et al., 2021), which could result in EAL learners acquiring the vocabulary in a different order than ESL or EFL learners. The results of our study, which support previous research, indicate a notable disparity in vocabulary knowledge growth, particularly within the 3000-word band. In this band our participants demonstrated a mastery level that was noticeably lower (by more than 10%) compared to the mastery they displayed in the 4000-word band. Table 9 highlights this pattern with examples from individual learners. The misalignment between the expected and actual vocabulary acquisition sequences, where learners are expected to acquire bands in order, underscores the challenges associated with using the BNC/COCA for teaching mid- to low-frequency vocabulary to this particular group of learners.

Furthermore, given that previous studies have shown that vocabulary acquisition takes place at a relatively steady pace (Dixon et al., 2020; Webb & Chang, 2012), it is unlikely that learners would be able to acquire the vocabulary knowledge required to gain the 95% to 98% coverage necessary for comprehension during their time at school using the BNC/COCA alone. This supports calls from researchers (Coxhead & Boutorwick, 2018;

Nation, 2016) that highlight the need to develop domain-specific word lists for EAL learners in order to help them acquire the vocabulary knowledge they need to succeed academically.

8. Implications and Limitations

The current study has several limitations that should be considered when interpreting its findings. One potential issue, as previously discussed, is the use of two different vocabulary assessment tools to measure the vocabulary knowledge of the participants in the study. In the future, it would be preferable to conduct a similar study with the same types of assessment for both groups of learners. However, it is important to keep in mind that both assessment tools used in this study were developed for adult EFL learners, and neither may be entirely suitable for EAL learners. Unfortunately, there are limited assessment tools for use with this group of learners. While tools do exist for younger EAL learners, such as the Peabody Picture Vocabulary Test (Dunn & Dunn, 2012), developed for young FLE speakers, and the English Picture Vocabulary Test (Güngör & Önder, 2023), developed for young L2 learners of English, there are no tools that have been specifically designed to measure the vocabulary knowledge for the age group of EAL learners participating in this study. The development of such a test is important, based on both the findings of this study as well as those of previous research (Green & Lambert, 2018; Greene & Coxhead, 2015), which have shown that the vocabulary needed in the EMI classroom is different from the vocabulary found in frequency lists developed for adult learners. A further limitation of the current study relates to the content of the corpus. The corpus used for this study was composed entirely of textbooks developed for the classes being taught. While it would have been beneficial to include a spoken component in the corpus, the logistics of making recordings in the classroom and transcribing a sufficient number of those recordings to develop a spoken corpus of a comparable size to the written corpus made it impossible to undertake as part of this study. Given the acknowledged differences between spoken and written English (Dang et al., 2017), we feel that it would be beneficial to conduct a follow-up study using a corpus of spoken English. A final limitation that needs to be addressed is the effect of L1 backgrounds, which the literature has shown can affect learners' English vocabulary knowledge (Booth & Clenton, 2020). While our dataset includes assessments from learners with diverse linguistic backgrounds, we were unable to evaluate how these backgrounds might impact their knowledge of academic English. We would like to pursue this in the future should we be able to include a sufficient number of learners from similar L1 backgrounds. We would also like to expand the scope of our study to include EAL learners studying in schools outside of Japan.

Despite the limitations discussed above, the current study offers some important implications with regard to supporting EAL learners in the classroom. First, as suggested by previous studies (Coxhead & White, 2012; Green & Lambert, 2019; Greene & Coxhead, 2015), educators looking to support EAL should prioritise enhancing their students' vocabulary knowledge. The data from this study highlight a pressing need to enable learners to expand their vocabulary knowledge in order to comprehend the textbooks they are asked to read. Additionally, there is an urgent requirement to develop word lists specifically tailored for EAL learners. While lists such as the AWL (Coxhead, 2000) and the BNC/COCA (Nation, 2020) are able to provide EAL learners studying in the international school context some support in acquiring the vocabulary they need to understand the textbooks that they are being asked to read, they are often either too broad (in the case of the BNC/COCA) or too limited (in the case of the AWL) in scope to provide learners with the support they require. While including important words, tools like the AWL were developed using a corpus of articles written for adult learners and may not include all of the words that EAL learners need in their context. On the other hand, the BNC/COCA contains too many

words, requiring EAL learners to acquire upwards of 7000 new words before they are able to achieve sufficient coverage to understand the textbooks from subjects such as biology or chemistry.

Furthermore, both sets of word lists, by design, cover vocabulary from multiple domains, making it difficult for teachers to select the words most appropriate for a single subject. This is important because research shows us (e.g., Teng, 2019; van Zeeland, 2013) that it is more effective to teach words in context. A set of domain-specific word lists focusing on the various subjects that EAL learners studying in the international school context would be likely to study would allow teachers to focus on a manageable list of words and teach them within the context of the subject where those words are required.

9. Conclusions

Our research emphasises the vocabulary gaps among EAL learners, highlighting the need for targeted vocabulary support in the classroom. The crucial role of vocabulary in EAL learners' reading comprehension shown by previous studies (e.g., Brooks et al., 2021; Marianne & Coxhead, 2023), along with the gaps in this knowledge illustrated by the current study, further underscores the need to prioritise vocabulary acquisition. Equipping EAL learners with the academic vocabulary they need can be achieved through classroom instruction and additional support outside of the classroom. This study highlights why placing greater emphasis on vocabulary instruction is vital for EAL learners to effectively comprehend the texts they encounter in their coursework.

Nonetheless, viewing vocabulary in the broader context of the types of texts EAL learners are being asked to read for their class is also essential. Therefore, teachers within the EAL setting, where content learning often takes precedence over language acquisition, need to be given the tools to identify the most relevant vocabulary for their learners. This will allow them to integrate activities that enhance the vocabulary knowledge and overall English language skills that learners need for the classes that they are enrolled in. We hope that this study will provide a starting point for such research in the future.

Author Contributions: Conceptualization, G.B., J.C. and S.F.; Data curation, G.B. and J.C.; Formal analysis, G.B.; Funding acquisition, G.B., J.C. and S.F.; Investigation, G.B. and J.C.; Methodology, G.B., J.C. and S.F.; Project administration, G.B., J.C. and S.F.; Resources, G.B., J.C. and S.F.; Supervision, J.C.; Validation, G.B.; Visualization, G.B.; Writing—original draft, G.B.; Writing—review & editing, G.B., J.C. and S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by two Grants-in-Aid for Scientific Research (No. 17K03035 and No. 20K00793) from the Japan Society for the Promotion of Science. The authors are very grateful for this support.

Institutional Review Board Statement: The study has been approved by the Research Ethics Committee of the Graduate School of Humanities and Social Sciences, Hiroshima University (approval number: HR-HUM-000762).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We extend our sincere gratitude to the two anonymous reviewers and the *Languages* production team for their valuable comments and suggestions. Their insightful feedback has greatly improved the quality of our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Afitska, O., & Heaton, T. J. (2019). Mitigating the effect of language in the assessment of science: A study of English-language learners in primary classrooms in the United Kingdom. *Science Education*, *103*(6), 1396–1422. [CrossRef]

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice: A Publication of the Division for Learning Disabilities, Council for Exceptional Children*, *20*(1), 50–57. [CrossRef]

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

Benjamin, R. G. (2011). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88. [CrossRef]

Benson, S., & Madarbakus-Ring, N. (2021). A comparison of textbook vocabulary load analysis. *Vocabulary Learning and Instruction*, *10*(2), 9–17. [CrossRef]

Booth, P., & Clenton, J. (2020). *First language influences on multilingual lexicons*. Routledge. [CrossRef]

Brooks, G. (2023). *Bridging the vocabulary gap for English as an additional language learners* [Ph. D. thesis, Hiroshima University].

Brooks, G., Clenton, J., & Fraser, S. (2021). Exploring the importance of vocabulary for English as an additional language learners' reading comprehension. *Studies in Second Language Learning*, *11*(3), 351–376. [CrossRef]

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology*, *58*(5), 412–424. [CrossRef]

Carder, M. (2007). Bilingualism in international schools. *Multilingual Matters*. [CrossRef]

Castellano-Risco, I., Alejo-González, R., & Piquer-Píriz, A. M. (2020). The development of receptive vocabulary in CLIL vs EFL: Is the learning context the main variable? *System*, *91*, 102263. [CrossRef]

Chuang, H., Joshi, R. M., & Dixon, L. Q. (2012). Cross-language transfer of reading ability. *Journal of Literacy Research*, *44*(1), 97–119. [CrossRef]

Clegg, J., & Afitska, O. (2011). Teaching and learning in two languages in African classrooms. *Comparative Education Review*, *47*(1), 61–77. [CrossRef]

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

Coxhead, A. (2012). Researching vocabulary in secondary school English texts: 'The hunger games' and more. *English in Aotearoa*, *78*, 34–41.

Coxhead, A. (2017). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives* (Routledge research in English for specific purposes). Routledge. [CrossRef]

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science. *TESOL Quarterly*, *52*(3), 588–610. [CrossRef]

Coxhead, A., Demecheleer, M., & McLaughlin, E. (2016). The technical vocabulary of Carpentry: Loads, lists and bearings. *TESOLANZ Journal*, *24*, 38–71.

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121–135. [CrossRef]

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, *16*(2), 37–52.

Coxhead, A., & White, R. (2012). Building a corpus of secondary school texts: First you have to catch the rabbit. *New Zealand Studies in Applied Linguistics*, *18*(2), 67–73.

Dale, E., & O'Rourke, J. P. (1976). *The living word vocabulary: The words we know, A national vocabulary inventory*. Field Enterprises Educational Corp.

Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge. [CrossRef]

Dang, T. N. Y., Coxhead, A., & Webb, S. A. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997. [CrossRef]

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics (Association for Computational Linguistics)*, *42*(3), 457–490. [CrossRef]

Dixon, C., Thomson, J., & Fricke, S. (2020). Evaluation of an explicit vocabulary teaching intervention for children learning English as an additional language in primary school. *Child Language Teaching and Therapy*, *36*(2), 91–108. [CrossRef]

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first-and second-language learners. *Reading Research Quarterly*, *38*(1), 78–103. [CrossRef]

Dunn, L. M., & Dunn, L. M. (2012). *Peabody picture vocabulary test* (3rd ed.). American Guidance Service: PsycTESTS Dataset. APA PsycTests. [CrossRef]

Faitaki, F., Hessel, A., & Murphy, V. A. (2022). Vocabulary and grammar development in young learners of English as an additional language. In M. Schwartz (Ed.), *Handbook of early language education* (pp. 428–444). Springer International Publishing. [CrossRef]

Flor, M., Holtzman, S., Deane, P., & Bejar, I. (2024). Mapping of American English vocabulary by grade levels. *ITL—International Journal of Applied Linguistics*, *175*(1), 25–45. [CrossRef]

Fraser, S. (2010). *The lexis of pharmacology texts: A corpus linguistic analysis* [Ph.D. thesis, Swansea University].

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. [CrossRef]

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes*, *35*, 105–115. [CrossRef]

Green, C., & Lambert, J. (2019). Position vectors, homologous chromosomes and gamma rays: Promoting disciplinary literacy through secondary phrase lists. *English for Specific Purposes*, *53*, 1–12. [CrossRef]

Greene, J. W., & Coxhead, A. (2015). *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*. Paul H. Brookes Publishing.

Groves, F. H. (2016). A longitudinal study of middle and secondary level science textbook vocabulary loads. *School Science and Mathematics*, *116*(6), 320–325. [CrossRef]

Güngör, B., & Önder, A. (2023). Development of English picture vocabulary test as an assessment tool for very young EFL learners' receptive and expressive language skills. *Early Education and Development*, *34*(2), 572–589. [CrossRef]

Ha, H. T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia*, *11*(1), 20. [CrossRef]

Hancke, J., Vajjala, S., & Meurers, W. D. (2012, December 8–15). *Readability classification for German using lexical, syntactic, and morphological features*. International Conference on Computational Linguistics (pp. 1063–1080), Mumbai, India.

Hessel, A. K., & Murphy, V. A. (2019). Understanding how time flies and what it means to be on cloud nine: English as an Additional Language (EAL) learners' metaphor comprehension. *Journal of Child Language*, *46*(2), 265–291. [CrossRef]

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, J., Gao, X., & Qiu, X. (2021). Lexical coverage and readability of science textbooks for English-medium instruction secondary schools in Hong Kong. *SAGE Open*, *11*(1), 215824402110018. [CrossRef]

International Baccalaureate. (2023). *International baccalaureate facts and figures*. Available online: https://www.ibo.org/about-the-ib/facts-and-figures/ (accessed on 6 September 2023).

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. [CrossRef]

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In C. Lauren, & M. Nordman (Eds.), *From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307–322. [CrossRef]

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

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

Leung, C. (2014). Researching language and communication in schooling. *Linguistics and Education*, *26*, 136–144. [CrossRef]

Luxton, J., Fry, J., & Coxhead, A. (2017). Exploring the knowledge and development of academic English vocabulary of students in New Zealand secondary schools. *Set: Research Information for Teachers*, *1*, 12–22. [CrossRef]

Marianne, & Coxhead, A. (2023). Getting to know your learners in an EAL context. In *EAL research for the classroom* (pp. 185–208). Routledge. [CrossRef]

Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities. *Reading in a Foreign Language*, *22*, 56–70.

McLean, S., & Kramer, B. (2016). The creation of a new vocabulary levels test. *Shiken*, *19*(1), 1–9.

Melby-Lervåg, M., & Lervåg, A. (2014). Effects of educational interventions targeting reading comprehension and underlying components. *Child Development Perspectives*, *8*(2), 96–100. [CrossRef]

Melby-Lervåg, M., Lyster, S. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. [CrossRef] [PubMed]

Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, *20*(1), 30–53. [CrossRef]

Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford University Press.

Murphy, V. A., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. Educational Endowment Foundation.

NALDIC. (2015). *EAL achievement: The latest information on how well EAL learners do in standardised assessments compared to all students*. NALDIC.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.
Nation, P. (2005). *Teaching and learning vocabulary*. Taylor Francis.
Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. [CrossRef]
Nation, P. (2020). *The BNC/COCA word family lists*. Available online: https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf (accessed on 14 August 2022).
Nguyen, C. (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. [CrossRef]
Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*(2), 189–208. [CrossRef]
O'Loughlin, R. (2012). Tuning In to vocabulary frequency in coursebooks. *RELC Journal*, *43*(2), 255–269. [CrossRef]
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
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. [CrossRef]
Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. [CrossRef]
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. [CrossRef]
Sharples, R. (2021). Teaching EAL: Evidence-based strategies for the classroom and school. *Multilingual Matters*. [CrossRef]
Smith, B. (2015). *US, UK set to lose market share of mobile students in next decade*. Available online: https://thepienews.com/news/us-uk-set-to-lose-market-share-of-mobile-students-in-next-decade/ (accessed on 21 February 2020).
Spencer, S., Clegg, J., Stackhouse, J., & Rush, R. (2017). Contribution of spoken language and socio-economic background to adolescents' educational achievement at age 16 years. *International Journal of Language & Communication Disorders/Royal College of Speech & Language Therapists*, *52*(2), 184–196. [CrossRef]
Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 102279. [CrossRef]
Taylor, S. E., Frackenpohl, H., & White, C. E. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Steck-Vaughn Company.
Teng, F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *Language Learning Journal*, *47*(2), 145–158. [CrossRef]
Townsend, D., Bear, D., Templeton, S., & Burton, A. (2016). The implications of adolescents' academic word knowledge for achievement and instruction. *Reading Psychology*, *37*(8), 1119–1148. [CrossRef]
Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*(5), 453–466. [CrossRef]
Uemura, T., & Ishikawa, S. (2004). JACET 8000 and Asia TEFL vocabulary initiative. *Journal of Asia TEFL*, *1*(1), 333–347.
US Department of Education, National Center for Education Statistics. (2022). *English learners in public schools*. Available online: https://nces.ed.gov/programs/coe/indicator/cgf/english-learners (accessed on 13 November 2022).
van Zeeland, H. (2013). L2 vocabulary knowledge in and out of context. *Australian Review of Applied Linguistics*, *36*(1), 52–70. [CrossRef]
Webb, S. A., & Chang, A. C. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126. [CrossRef]
Webb, S. A., & Nation, P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, *16*, 1–10.
Webb, S. A., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *International Journal of Applied Linguistics*, *168*(1), 33–69. [CrossRef]
West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.
Woodward-Kron, R. (2008). More than just jargon—The nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes*, *7*(4), 234–249. [CrossRef]
Xodabande, I., & Hashemi, M. R. (2023). Learning English with electronic textbooks on mobile devices: Impacts on university students' vocabulary development. *Education and Information Technologies*, *28*(2), 1587–1611. [CrossRef] [PubMed]
Yang, L., & Coxhead, A. (2022). A corpus-based study of vocabulary in the New Concept English textbook series. *RELC Journal*, *53*(3), 597–611. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

languages

MDPI

Article

Delivering an ESP Pedagogic Word List: Integrating Corpus Analysis, Materials Design, and Software Development

Simon Fraser *, Marshall Kiyoshi Higa and Walter Davies

Institute for Foreign Language Research and Education, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Japan; mhiga@hiroshima-u.ac.jp (M.K.H.); wdavies@hiroshima-u.ac.jp (W.D.)
* Correspondence: fraser@hiroshima-u.ac.jp

Abstract: With vocabulary playing an essential role in the learning of English for Specific Purposes, teachers face the challenge of organising and teaching lexis in a way that maximises opportunities for acquisition. Specialised word lists offer a solution, but a major obstacle is how to integrate these lists into learning materials containing items used in actual discourse. In this paper, we report on research involving the creation of a medical English word list (MEWL), integrated into a set of specially designed materials for students at a national university in Japan. These materials, developed through needs analysis at the university's medical school, are primarily organised around body systems, with an additional focus on doctor–patient communication. The MEWL is complemented by a list of word parts, aiming to sensitise students to complex medical terms. We describe the delivery of the list, first through the courses and materials, and then via the development of a vocabulary learning tool, Hi-Lex, which analyses texts against any word lists it contains. Hi-Lex allows learners to create personalised word lists and understand word usage in context. The findings of a small trial study of Hi-Lex (N = 31) illustrate how the software provides insight into students' selections of words in specialised texts.

Keywords: medical English education; materials development; corpus analysis; word list; vocabulary learning software

1. Introduction

With significant advances in software and computing over the past decades, the analysis of vast amounts of digitally stored text has become possible, resulting in increasingly sophisticated studies of vocabulary and language structure. One outcome of this has been the proliferation of word lists for both general English and English for Specific Purposes (ESP). In ESP, such lists have become tailored to specific domains, for example, law, engineering, and—the focus of the present research—medicine. These lists have an important role to play in providing second language (L2) learners at the university level with the most relevant vocabulary in their chosen field. However, from a pedagogic viewpoint, a key question is how to create a word list and structure the learning of vocabulary in a way that is optimal for students. To address this concern, a research team at Hiroshima University has been working on a project that utilises needs analysis and corpus-assisted techniques to develop a specialised word list, the Medical English Word List (MEWL), which is integrated into a set of English course materials for undergraduate medical students. The aim of the project is to provide a syllabus, materials, and word lists that aid students in making the transition from general English to medical English. The syllabus is a quasi-parallel one, primarily focused on the core content that they are already studying

in their medical courses but going slightly beyond it where the complexity of the English is well within their grasp.

Of particular interest to the research team is how vocabulary acquisition fits into the learning process within the project and how a better understanding of this will inform medical English curriculum design as well as teaching strategies. In this context, Nation (2024) draws our attention to eight principles that can be applied to the teaching and learning of vocabulary (focus, accuracy, repetition, time-on-task, elaboration, analysis, motivation, and self-efficacy). By applying these principles and tailoring approaches to align with the ways in which learners naturally acquire words, the effectiveness of instruction can be maximised. To this end, we have incorporated Hi-Lex, a vocabulary-learning tool developed by a member of the research team. The software makes use of spaced repetition, allowing learners to analyse texts of their choice and focus on the words they feel are most useful.

The aim of this paper is twofold: first, to document the development of medical English materials from which the MEWL emerges, and second, to investigate which words students focus on when they record terms in Hi-Lex. Given that the software is still under development, spaced repetition learning is not investigated in this study. However, we plan to conduct future research to explore its effectiveness and integration once the software is fully operational.

We begin by summarising the developments in ESP that provide the background to the project and highlighting the important role played by vocabulary. We then detail the initial stages in the development of medical English materials in relation to needs analysis and syllabus design, explaining how we have integrated corpus analysis with materials building. This is followed by an outline of the approach used in the construction of the Medical English Word List (MEWL). In the final section of the paper, we present a small trial study of the Hi-Lex software, demonstrating how learners can analyse content and focus on the words they wish to learn, rather than those chosen by the course designers and teachers.

While the paper is written in British English, the medical terms used throughout are presented in American English, reflecting the language in the units of material, which follows the American conventions more commonly used in Japan. This ensures consistency with the materials and accurately represents the language students encounter in their learning context.

2. Background

2.1. Developments in English for Specific Purposes (ESP)

Over the years, ESP has undergone major transformation to meet the demands of global communication, technological integration, and changing pedagogical methods. In this section, we outline some of the trends and developments that have significantly shaped the medical English project.

2.1.1. Technological Integration in ESP

The rapid development of technology has compelled ESP practitioners to swiftly adapt to AI-driven advancements, including software writing aids, vocabulary-learning tools, games, and automatic translation systems. These innovations have significantly impacted course provision and delivery, particularly regarding learning management systems like Moodle. As a result, flipped learning approaches have become increasingly prevalent. Flipped learning allows theoretical content to be explored outside the classroom, while class time is dedicated to interactive, practical learning (Bergmann & Sams, 2015). Several studies (e.g., Enokida et al., 2018; Hsiao et al., 2023; Liu, 2022) have investi-

gated the effectiveness of flipped learning as an instructional approach in ESP, highlighting its potential for promoting learner engagement and autonomy.

2.1.2. Needs Analysis and Relevance

In the field of ESP, there is a growing emphasis on conducting comprehensive needs analyses. This ensures that language instruction corresponds closely with learners' specific professional or academic areas. According to Anthony (2018), a comprehensive needs analysis involves not only learner surveys but also diagnostic, democratic, and analytic perspectives from a variety of stakeholders. Research by Hutchinson and Waters (1987), Belcher and Lukkarila (2011), Chan (2018), and Lu (2018) has examined the role of needs analysis in ESP course design, emphasising the importance of matching curriculum content with learners' specific language needs and goals. In medical English, Simpson's (2022) study is a good example of this, with the findings of his needs analysis indicating that curricula should contain opportunities for cognitive, motivational, and identity formation in addition to the actual English language required by medical professionals.

2.1.3. Corpus Linguistics in ESP

The application of corpus linguistics is enhancing the development of materials that address the precise lexical and grammatical demands of specific fields. Corpus-based software is used to identify relevant vocabulary items and usage patterns within the target domain, allowing researchers to analyse large sets of language data efficiently. One example is the development of word lists (see Section 2.2.2). Another is data-driven learning (DDL), as discussed by Johns (2002), which allows learners to explore language behaviour through corpus-analytic tools. Boulton et al. (2012) offer an overview of corpus use across diverse areas, including translation, specialised language research, and ESP material development. Their work demonstrates how corpus-based approaches inform language selection and teaching strategies. Flowerdew (2015) advocates corpus-based approaches to enrich vocabulary instruction with authentic language data. Corpora provide language samples from specialised domains, revealing high-frequency terms and common collocations relevant to specific fields. The use of such resources plays a particularly important role in facilitating vocabulary acquisition.

2.2. The Importance of Vocabulary in ESP

In ESP, the identification and categorisation of academic and disciplinary vocabulary is important for a range of learners, researchers, educators, and materials designers (Coxhead, 2018). For learners, mastery of the relevant vocabulary is vital if they are to effectively engage with their professional domains. Vocabulary knowledge equips them with the necessary tools to understand technical texts, communicate efficiently, and perform domain-specific tasks. Key studies highlight its significance: Coxhead (2000) and Gardner and Davies (2014) have compiled word lists to identify essential academic vocabulary across disciplines, while other researchers, such as Dang and Webb (2016), emphasise the importance of beginning with high-frequency vocabulary relevant to learners' needs. Meanwhile, Schmitt (2010) highlights the significance of formulaic sequences for fluent communication, while Ward (2007) explores the relationship between collocation and technicality. Additionally, in fields such as medicine and biology, which rely heavily on specialised terminology, an understanding of word formation and morphology is particularly important. The ease with which a word can be decomposed into its constituent morphemes has been termed its "derivational complexity" by Laufer (1997), meaning that the difficulties faced by learners will depend on their familiarity with the most frequently occurring affixes. Take *desensitization*, for example: if learners know *sensitive*, which occurs quite frequently when discussing medical conditions, and understand the morphemes *de-*,

-ize, and *-ation,* then they should be able to work out the meaning with relative ease. ESP courses may include exercises and tasks focused on derivational morphology to help learners understand the structure and derivation of specialised vocabulary (see, e.g., Chabner, 2023). By analysing word roots and combining forms (e.g., *oste/o-*, meaning 'bone') and affixes (e.g., *anti-*, meaning 'against'; *-itis,* meaning 'inflammation'), learners can deduce the meanings of unfamiliar terms and expand their lexicon more effectively. Also fundamental to the formation of medical terminology is compounding, in which two existing words are combined to form a specific meaning. With transparent compounds, the meaning can easily be inferred (e.g., *painkiller*), but a word like *heartburn*, an opaque compound, requires specialised knowledge to understand the actual medical condition.

Lexical studies have informed effective ESP pedagogical practices and curriculum development, helping to ensure that vocabulary instruction is tailored to meet the specific needs of learners in their professional or academic contexts. We now turn our attention to the characteristics of the different categories of specialised lexis and to the work being carried out to equip ESP learners with the most relevant vocabulary.

2.2.1. Categorising Specialised Vocabulary

The vocabulary of specialised texts is often divided into *technical* and either *subtechnical* or *semi-technical* categories (e.g., Cowan, 1974; Inman, 1978; Trimble, 1985; Farrell, 1990; Lam, 2001). It has also been classified according to four-step/category rating scales (Chung & Nation, 2003; Fraser, 2005). However, Fraser (2006, 2012) argues that these classifications lack accuracy and consistency in their usage across the literature and suggests instead adopting alternative categories that better capture the nuances of a specialised lexicon: *fully technical*, *cryptotechnical*, *lay-technical*, *academic*, and *general* words.

Fully technical terms are specific to their fields, and in areas such as science, medicine, and law, they are typically of Graeco-Latin origin, examples being *photosynthesis* (science) and *stenosis* (medicine). Such words are unfamiliar to general language users, who do not have specialised training or expertise in a particular field. Technical terms may also contain general words, as with *short head* (of a muscle) and *inner ear*. Cryptotechnical vocabulary comprises polysemous words, commonly found in general language but possessing an additional technical meaning (or meanings) that may elude non-specialists, such as *inferior*. These terms, although often labelled subtechnical or semi-technical—implying that they somehow lack 'technicalness'—function in a similar way to unambiguously technical words when encountered in domain-specific texts. Lay-technical words, while clearly technical across a variety of disciplines, retain a basic meaning understandable to those without specialist knowledge in the field. Additionally, academic words, such as *method* and *conclusion,* which are essential for scholarly discourse, along with general words that help to link ideas and are easily understood by a broad audience, play significant roles in specialised writing. By adopting these refined categories, a more accurate representation of the diverse lexical landscape found in specialised texts is achieved, reducing ambiguity and confusion while enhancing both comprehension and communication within specialised domains.

In summary, all categories of technical vocabulary, academic words, and general words are important for learners of ESP, but their relative importance may depend on factors such as the learners' specific field or profession, their academic or professional goals, and the requirements of their language learning setting. In the context of English for Medical Purposes, the main area of difficulty for Japanese learners is likely to arise from the differences between the two languages regarding highly specialised terms; while English relies on Latin and Greek roots, Japanese, although incorporating elements from these languages, often uses *kanji* (Chinese characters) to convey meaning. While many Japanese

medical terms are direct loanwords transliterated from Latin or Greek into the Japanese syllabary (*katakana*), pronunciation of these words can be a challenge, affecting understanding and communication. In a setting where precise language is crucial, this underlines the importance of context and clarity. Ideally, ESP instruction should incorporate a balanced approach that addresses both technical terminology and non-technical vocabulary to support learners' language development and proficiency within their specific field.

2.2.2. Pedagogical Word Lists

Numerous studies have been conducted with the aim of identifying and listing the most useful words for learners across a variety of fields. In EAP (English for Academic Purposes), lists such as the Academic Word List (AWL, Coxhead, 2000), the New Academic Word List (NAWL, Browne et al., 2013a), and the Academic Vocabulary List (AVL, Gardner & Davies, 2014) have been developed to provide learners with words that are frequent across a wide range of disciplines. Lists of key technical and sub-technical words have also been produced for specialised fields such as electronics (Farrell, 1990), engineering (Ward, 1999), and business (Browne & Culligan, 2015). Recent years have seen a proliferation of specialised lists in areas as diverse as plumbing (Coxhead & Demecheleer, 2018), English as an Additional Language (Brooks, 2023), and veterinary medicine (Özer & Akbaş, 2024). The medical field has its fair share of word lists, including those created by Fraser (2007, 2009), Hsu (2013), Lei and Liu (2016), Wang et al. (2008), and Yang (2015).

2.3. Challenges with Teaching Specialised Vocabulary and Using Word Lists

There is, as we have seen, no shortage of pedagogical word lists. Nevertheless, teaching specialised vocabulary and utilising word lists in ESP contexts present numerous challenges. The first of these is the need to ensure that the vocabulary taught closely matches the specific needs, context, and professional or academic domain of learners. Medicine, the focus of the current research, has plenty of word lists. However, lists such as the Medical Academic Word List (Wang et al., 2008) and Medical Academic Vocabulary List (Lei & Liu, 2016) tend to prioritise the academic vocabulary of medical research articles rather than, for example, the terms being learned by undergraduates at the early stages of their studies. For such learners, Quero and Coxhead (2018) argue that a focus on the high-frequency terms occurring in medical textbooks is essential if learners are to understand and produce medical discourse. The correct alignment of vocabulary and needs is key: classroom time is limited, and learners need to know the most useful words associated with their specific area of study. As well as being an inefficient use of time, the study of vocabulary that is irrelevant to the students' field and current level of learning is likely to decrease motivation, lead to frustration, and foster disengagement.

In a field such as medicine, challenges include the huge number of fully technical terms, which are complex in form, and cryptotechnical words, which are difficult for learners to grasp and retain without adequate context. This is not helped by the fact that available word lists often consist of a large number of decontextualised items. It is important to ensure that the vocabulary presented in word lists is linked to carefully constructed instructional materials if learners are to be able to understand and use it in context.

In addition, when designing materials and constructing word lists, it is necessary to consider how words are acquired, and what can be done to help learners in this process. Strategies for reinforcing vocabulary learning, such as spaced repetition and contextualised practice, are essential for long-term retention (Nation, 2013). Integrating vocabulary instruction with other language skills and designing valid and reliable assessment tasks also present challenges that require effective instructional strategies and ongoing evaluation.

To address these challenges, developing a comprehensive understanding of learners' needs and continually adapting teaching materials and approaches is important to optimise vocabulary instruction. The following sections of the paper focus on our efforts to achieve this in a medical English context.

3. The Medical English Project

3.1. Beginnings

This project was started by a research team from the Institute for Foreign Language Research and Education (FLaRE) at Hiroshima University, one of the national universities in Japan. In 2012, the dean of the medical school contacted the institute to ask if there was a way in which it could support medical English learning for students in their third year of medical studies. The result of the discussions was the creation of a small team of four experienced language teachers to provide a four-day intensive course to third-year students on the medical campus. Materials for the course were built from a variety of sources, including commercially produced CDs, DVDs, and the internet (see Davies et al., 2013). In addition, a short word list (or glossary) was created to supplement the materials. Student feedback was positive, particularly in relation to the word list, and the medical school asked the team to provide the course again as part of the medical English curriculum. At this point, two members of FLaRE (who are also authors of this paper) and its director formed a team to conduct research into medical English and develop a more comprehensive course. Although there have been changes over the years, the core members of the research team are all applied linguists from FLaRE, and two medical doctors have also acted in an advisory capacity throughout the project.

In the first year of research, medical professors at the university hospital were interviewed, and a meeting with senior medical staff was held to discuss ideas. The research initially resulted in the creation of seven units of material for the intensive course, and then, through collaboration with the medical school's English teacher, a further seven units were developed to create a unified curriculum. A specific aim of the project was to develop a word list embedded in the materials, allowing the terms to be learned through the usual tasks and activities undertaken by students in a communicative approach. Throughout the process, corpus analysis techniques were used to refine and develop the materials, with the word list emerging through the extraction of terms from the materials, and a word-part list being created through an investigation of those terms.

3.2. Situation, Resource, and Needs Analysis: The SiReN Model

In the initial stages of most ESP projects, research activities are often categorised under the label of "needs analysis". However, we found this term too broad to accurately reflect the decisions being made by the research team. As a result, we developed a three-concept model that considers situation and resources alongside needs (the SiReN approach, introduced in Fraser & Davies, 2022).

Situation analysis relates to the organisational and administrative structures within which the project takes place. In our case, the project was developed based on the provision of a course that was integrated into the credit-based structure of the university medical school's curriculum. With regard to English-language learning, in 2012, undergraduates studied general English in their first two years of university study. Subsequently, medical students focused on medical English for an additional year.

Resource analysis considers factors such as personnel, technology, time, and budget. Teachers and students had access to spacious, well-equipped classrooms with up-to-date audiovisual equipment. Also, the project team consisted of teacher-researchers, with sufficient time in their schedules and budget allocated to research and development activities.

Between them, the members of the team had experience in corpus analysis, word-list creation, and materials development. This was particularly important in terms of an initial plan of research, which could be adapted as the project developed. The primary challenge was developing a word list that seamlessly integrated into a course. The initial plan was linear: A corpus of medical articles would be analysed to create a word list of high value items, and from this a set of materials would be created that was appropriate for third-year students, based on the team's experience of teaching the course in 2012. A further resource that developed during the project was a strong working relationship with senior members of the medical school who could offer help and advice at key points in the project.

Needs analysis is perhaps the most complex of the three concepts. In our case, we have divided it into *formal needs analysis* and *informal needs analysis*. Formal needs analysis primarily involved interviews and meetings with senior medical staff. Informal needs analysis, which is also very important, was oriented towards students. The key difference is that whereas formal needs analysis is a highly planned and structured process, informal needs analysis occurs spontaneously through chance conversations and interactions.

3.2.1. Formal Needs Analysis

Initial interviews were opportunistic, involving interaction with a neurosurgeon at a local hospital to identify how a practising doctor used English. In this context, regarding the English needs of healthcare professionals in a country like Japan, Guest's (2022) distinction between *horizontal* (peer–peer, professional–professional) and *vertical* discourses (healthcare professional–patient/client) is useful. In the neurosurgeon's case, English use was primarily through horizontal discourses in the form of medical articles, with him dividing these into three main categories: basic research, clinical research, and case studies. His purpose was to keep up with developments in his field by reading articles both in Japanese and English.

The next set of interviews took place at the university medical school, where three senior medical professors were asked for their ideas on what was important for students to learn in English. Each of them stressed the importance of anatomy, with other suggestions being physiology, biochemistry, medical examinations, and common diseases. Regarding recommended texts on medicine, key books mentioned in the interviews were *Gray's Anatomy*, *Grant's Anatomy*, and *Harrison's Principles of Internal Medicine*.

A significant challenge identified in the needs analysis was how to balance the immediate educational requirements of a third-year medical student with the long-term professional demands of a medical practitioner. In terms of horizontal discourses, practitioners need to excel in professional communication, which includes making presentations, reading, and writing research articles in English; for vertical discourses, they might have to treat patients in English, involving cross-cultural communication. However, medical education spans a lengthy period, and third-year students are still at the stage where they are engaged primarily in classroom-based learning, focusing on medical terminology, examination preparation, and clinical skills development.

In a meeting to discuss research and development, a group of six senior medical professors raised issues regarding the team's original strategy, which focused on medical research articles. They pointed out that such an approach might not adequately address the practical learning needs of third-year students. This led us to reassess how corpus analysis might best be deployed, resulting in a decision to focus on two reference books whose importance had been highlighted in the interviews: *Gray's Anatomy for Students* and *Harrison's Principles of Internal Medicine*. However, one very important insight into materials development arose from the initial research article-based plan, which required the division of the medical field into ten areas. Such an approach ensures a broad and comprehensive

coverage of the medical field while maintaining a manageable level of segmentation for effective language instruction. This framework of ten areas, based on the suggestions of a senior professor, would guide the project when combined with the results from the informal needs analysis.

3.2.2. Informal Needs Analysis

Information from students was collected more informally through opportunistic discussions with second-year medical undergraduates enrolled in general English courses. In response to the emphasis placed on anatomy by the medical professors, one member of the project team set aside a small portion of his classes with students on the medical campus to introduce anatomical terms. This was achieved through tasks matching vocabulary to diagrams, leading to discussions with students in which they described the content of some of the books they were studying. Their anatomy/dissection textbook was written in Japanese, with key terms in English tagged to diagrams and added to the discourse. In addition, students often studied English medical terms through a series of books called *Kikutan*, which lists terms alongside contextualising sentence examples together with translations. For the detailed study of anatomy, they typically used *Netter's Atlas of Human Anatomy*, effectively a set of diagrams of human anatomy that map the body. In the Japanese version, diagrams are labelled in both Japanese and English, providing students with exposure to numerous English terms. However, these were not encountered in extended passages of text.

From the perspective of the research team, a breakthrough occurred when comparing the list of ten medical areas from the formal needs analysis with the anatomy diagrams. The link between the two was body systems. Although this was not a perfect match (for example, oncology, the study of cancers, is not confined to any single system), most fields could be matched with a body system, with the original list being edited and extended to thirteen fields (Table 1).

Table 1. Linking medical fields to body systems.

Medical Field	Body System
Neurosurgery	Central Nervous System (Brain)
Orthopaedics	Musculoskeletal System
Cardiology	Circulatory System (Heart)
Pulmonology	Pulmonary System
Gastroenterology	Digestive System
Endocrinology	Endocrine System
Hepatology	Liver
Immunology	Lymphatic System
Dermatology	Integumentary System
Nephrology/Urology	Urinary System
Gynaecology/Urology	Reproductive System
Otorhinolaryngology	Ear, Nose, and Throat
Ophthalmology	Eye

4. Initial Materials: Discourse, Vocabulary, and Task

4.1. Integrating Corpus Analysis and Materials Development

With the shift in corpus analysis from research articles to reference books, a new plan was developed. Corpus analysis and materials development would operate in parallel,

with analysis of the key medical texts serving as a guide to revise and enrich the materials, particularly in relation to important terms. Intuitively, it seemed easier to extract words from enriched materials than to build materials from an initial word list. This was a primary reason for creating a new word list instead of relying on existing ones. The list would be derived from the key medical fields that students were studying in Japanese, enabling them to connect the terms to their understanding of these fields in their first language. Also, this would help to control the number of words. With materials as the main focal point, we anticipated that a word list tailored to the learners' needs, and of manageable length, would emerge.

4.2. Embedding Key Vocabulary and Phrases Within Discourse

The aim of the research team was to construct materials that aligned broadly with a communicative approach to teaching and learning. Our plan included addressing communication in two key registers: doctor–doctor discourse and doctor–patient discourse. The needs analysis had indicated that although students had access to medical English terms, the books they were studying did not provide longer stretches of language. In addition, there was the challenge of linking the two registers. The emerging ideas in the syllabus were content-based, with a focus on body systems, meaning that passages of discourse in the form of essays on medical content and dialogue relating to doctor–patient interaction could be constructed. From a pedagogic perspective, the most straightforward doctor–patient interaction for linking the two registers would be the process of taking a patient's medical history, which involves interviewing the patient to make an initial assessment. This approach aligns with the insights of Wette and Hawken (2016), who emphasise the importance of developing clinical communication skills that extend beyond mere linguistic competence. Example dialogues and fictional patient profiles could be created for use in role plays, with each fictional patient presenting a different medical issue related to the body system under study.

Beginning with the central nervous system (CNS), and utilising the advice of a neurosurgeon, the following four medical conditions were chosen: *chronic subdural hematoma*, *subarachnoid hemorrhage*, *brain tumor*, and *cerebral infarction*. A history-taking dialogue was constructed as an example of doctor–patient interaction, along with fictional patient information for role plays. An essay on neurosurgery was composed, incorporating information about the four medical problems, while another essay on anatomy and physiology was written, focusing on key anatomical terms. In the case of the CNS, there was a strong focus on the lobes of the *cerebrum* and the layers of bone and tissue surrounding it, particularly the *meninges* (the *dura mater*, *arachnoid membrane*, and *pia mater*). Pedagogically, the unit would build from anatomy/physiology into descriptions of medical problems, then into doctor–patient discourse. A final section on treatments was added. A range of exercises and communicative tasks were developed around the discourse, with some focused on vocabulary building and others designed for communicative practice. Initially, six units of material were created in this format. Following the completion of these units, students were taught the content in an intensive course. Student evaluation of the classes was positive, and the course was repeated over a period of years. As research progressed, the team started construction of an extra unit for their materials (see Section 4.3 below) and a further seven units for the medical school's English instructor, resulting in the creation of an integrated syllabus of 14 units containing a large number of medical terms. Vocabulary tests formed part of the evaluation for each group of seven units, providing a stimulus for students to review and learn vocabulary that they had studied.

4.3. The Role of Corpus Analysis

Using earlier versions of *Antconc* (now updated to Version 4.3.1; Anthony, 2024), we created two corpora: one from *Gray's Anatomy for Students* and one from *Harrison's Principles of Internal Medicine* (for details of corpus construction, see Fraser et al., 2014, 2016). The purpose of these was not to create a pedagogical word list, but to provide tools for building teaching units that incorporate the most useful words and expressions, as determined by their frequency, keyness, and distribution across the different sections of the reference books. Through concordance analysis within *AntConc*, we were able to identify major gaps in the medical materials, refine the language used, enhance the appropriateness of collocations, and add important terms, thereby significantly improving the overall quality of the content.

One major gap in the materials related to technical positional terms. In analysing a list of single-word items from *Gray's Anatomy for Students*, a variety of adjectives stood out that included *superior*, *inferior*, *anterior*, and *posterior*. These terms were clearly important, although they did not fit neatly into any particular body system. A review of the chapters in the book indicated to us that a unit on the anatomical position (the standard body position for reference in medicine), positional terms, and views of the body needed to be created.

One example of refining the language in the materials was using a more appropriate verb to re-write the sentence "the blood *passes* into other arteries" as "the blood *flows* into other arteries" in the cardiovascular unit. An example of enriching the materials involved adding a sentence that incorporates the key terms *systole* and *diastole*.

5. Development of the Medical English Word List (MEWL)

A primary aim of the research was to develop a pedagogical word list that could serve both as a checklist and as a tool for acquiring new vocabulary. Importantly, unlike other available word lists, the MEWL, which consists of terms embedded in a set of specially designed course materials, fits the specific needs of the particular group of learners we are interested in. In this section, we outline our innovative approach to creating the list, which has been a significant objective alongside materials development. We describe the process of extracting the word list from the materials and detail the creation of a word-part list designed to aid students both in memorising encountered words and deciphering unfamiliar ones.

5.1. Extracting Words from the Materials

The final pedagogical word list was constructed from an analysis of the 14 units, with key terms being identified manually and listed on Excel sheets. For the initial selection of words, statistical software tools were not used. The team used its experience in teaching medical students, consultation with professors from the medical school, and the expertise of one of the authors, who has a background in a closely related discipline, pharmacology. Terms were chosen by carefully reviewing the essays and diagrams in the materials and choosing words that were deemed important.

Selection of a word or multiword term for inclusion was made on a case-by-case basis, with individual word types listed for their pedagogic usefulness (e.g., *vomit* and *vomiting* both occur frequently in the materials and are listed separately; *abnormality* is not found in the units but was added to *abnormal*, which does appear). To verify our selections, we examined words that were frequent (occurring several times in a particular chapter) and those that were widely distributed across the corpora of *Harrison's Principles of Internal Medicine* and *Gray's Anatomy for Students*, allowing us to identify important terms that may have been overlooked. A high frequency of occurrence, however, is not necessarily an

indicator of a term's relative importance. For example, *dura mater* appears 47 times in the anatomy corpus, while *pia mater*, which plays an equally vital role in the central nervous system, occurs only nine times.

The average number of terms per unit was 167, with a standard deviation of 27. However, a number of terms were duplicated across the sheets. The terms were collated onto one spreadsheet and, with duplications removed, the final list comprised just over 1700 items for analysis. Here, Fraser's (2006) categories could be applied, dividing terms into fully technical, lay-technical, and cryptotechnical vocabulary. Fully technical terms often involve Latin or Greek words or word parts, an example being *chronic subdural hematoma*. Lay-technical terms, those which are known to the nonspecialist, include *heart* and *knee*. Cryptotechnical terms are those which have a specialised meaning in medicine that differs from their ordinary use in everyday language. Examples of these are *superior* and *inferior,* which are purely locational terms in medicine with no reference to quality. Most of the terms in the list are single-word entries; however, there are also some that consist of two, three, or even four words (e.g., *selective internal radiation therapy*).

5.2. What Is the Nature of the Words in the List?

The majority of terms in the list are fully technical, and the research team has dedicated time to a thorough analysis of these. One of the advantages of setting up units that build from anatomy/physiology into medical problems is the connection between the words in a unit. For example, the relationship between ***dura*** *mater*—a tough membrane that surrounds the brain and spinal cord—and *sub**dura**l hematoma,* which is a collection of blood beneath the dura mater, can be easily illustrated (with the shared terms shown in bold). Fully technical terms can be considered to fall roughly into four categories: indivisible single words such as *bile*, divisible single words such as *astro/cyte*, multiword terms where the component words can be broken down as with *sub/arachnoid hemo/rrhage*, and multiword terms where the words are indivisible (e.g., *sweat gland*). In addition to the fully technical terms, there were lay-technical words such as *thumb*, *thigh*, and *wheeze,* and cryptotechnical terms like *posterior, swelling,* and *cavity.* Finally, there were a small number of colloquial equivalents of technical terms, such as *Adam's apple* ('laryngeal prominence'), *belly button* ('umbilicus'), and *throwing up* ('vomiting' or 'emesis').

To facilitate the learning of multipart and divisible terms, we conducted an analysis to identify and extract their components (roots, combining forms, affixes, etc.). The resulting list of word parts consists of 170 items.

With both the word list and the word-part list, items were identified and selected for inclusion from a practical pedagogic perspective. For example, the term *abdominal* is listed alongside *abdominal aorta* and *abdominal cavity*. The reason for this is that the latter two are important compound nouns in anatomy. However, *abdominal* itself collocates with a variety of other words such as *pain*, *surgery*, and *muscles*. When we looked at word parts, a similar issue emerged with the listing of *-ia*, *-emia*, and *-uria.* The suffix *-ia*, meaning 'condition', is valuable to teach on its own, but we considered *-emia* and *-uria* to be more easily learned as chunks, especially in terms such as *hyperglycemia* and *hematuria.* As we have noted, both the word list and word-part list serve as pedagogical tools, enabling students to familiarise themselves with medical English as they make the transition from general English to ESP.

6. Delivery and Software

Delivery of the word list comes through traditional classroom teaching and two software packages: the university's learning management system (LMS), and Hi-Lex, a vocabulary learning tool designed by one of the authors of this paper.

6.1. Combining Traditional Classroom and Online Learning

Initially, seven units of material taught by FLaRE members were made available online via the university's LMS (originally Blackboard Learn 9, subsequently Moodle) and used for flipped learning. Due to the COVID-19 pandemic, which made classroom instruction difficult, a further seven units were placed online and added to the existing materials. As a result, the initial learning of vocabulary items has been through online access to vocabulary exercises for essays from which the words are mainly drawn. Instilling an awareness in the students of the ways in which terms are built up of component parts has been achieved through classroom instruction of the seven units taught by the institute team. Work is currently underway on incorporating a word-part component into the remaining seven units.

Despite our success in delivering key medical terms and sensitising students to word parts, two important challenges remain. The first of these is that decisions regarding which words to focus on have been made by the materials designers, not by the students who study the material. The vocabulary exercises cover only a limited number of lexical items, and it may be that students want to learn terms from their study materials that the writers have not specifically highlighted. In addition, with the pressure of a demanding curriculum, resulting in students prioritising their core medical studies in Japanese over learning medical English, there is a need to provide them with an efficient study method that facilitates vocabulary retention. To address these issues, we decided to use Hi-Lex, a vocabulary learning tool developed at Hiroshima University (Higa & Ashida, 2023), which combines two functions: vocabulary analysis with reference to word lists, and spaced repetition learning. The software was initially designed as a combined learning and research tool for general English studies, but it can be adapted for use in ESP. In its current form it is not available for general use because of the server demands it requires; however, we hope that it will be accessible for broader use in the future.

6.2. Hi-Lex

Hi-Lex serves as a versatile educational tool that combines text analysis and word management functionalities, making it particularly useful for students aiming to enhance their vocabulary acquisition. Examples of more advanced text analysis tools include Laurence Anthony's *AntConc 4.3.1* (https://www.laurenceanthony.net/software/antconc, accessed on 14 February 2025), Tom Cobb's *Compleat Lexical Tutor 8.5* (https://www.lextutor.ca, accessed on 14 February 2025), and Atsushi Mizumoto's *New Word Level Checker* (https://nwlc.pythonanywhere.com, accessed on 14 February 2025). However, these tools are primarily aimed at academics rather than university students.

The word management list is similar to popular web applications such as *Anki* (https://apps.ankiweb.net, accessed on 14 February 2025), *Duolingo* (https://www.duolingo.com, accessed on 14 February 2025), and *Memrise* (https://www.memrise.com, accessed on 14 February 2025). These applications function like flashcards for vocabulary study and utilise spaced-repetition algorithms to determine review intervals, popularised by Piotr Wozniak's *SuperMemo* method (https://www.supermemo.com/en/supermemo-method, accessed on 14 February 2025).

6.2.1. Benefits of Hi-Lex: Integrating Text Analysis and Spaced Repetition

The Hi-Lex tool offers several key advantages, including the integration of text analysis and spaced repetition software into a single application, which reduces the steps students must take to begin learning words they consider important. Its text analysis features, though simple, allow students to select words based on individual profiles rather than relying on prescribed lists, a common practice in Japan. Additionally, although many spaced

repetition software options exist, most are designed for commercial use, making it difficult to access data on user word selection, frequency of use, and content for research purposes. Hi-Lex addresses this gap by providing researchers with valuable access to such information.

6.2.2. The Hi-Lex System

The Hi-Lex software has a workspace in which students can paste stretches of text and store them. Within the system, there is a set of word lists, and users can check whether a term appears in a particular list when deciding whether to choose it for learning in a spaced repetition software function. In addition, the texts themselves can be stored for future review. The ways in which the software can be used by students fall into two main categories, broadly equivalent to Nation's (2024) concepts of elaboration and analysis: it can help students consolidate the vocabulary covered in their courses prior to taking tests, and it can assist them in analysing the medical discourse that they choose to study or need to engage with outside medical English classes. For researchers and language instructors, it can be used as a tool for researching students' choices of words for learning.

The software is designed for Japanese users. To use Hi-Lex, the learner begins by copying and pasting text into the browser's workspace (ワークスペース) and then pressing the blue button (see Figure 1).

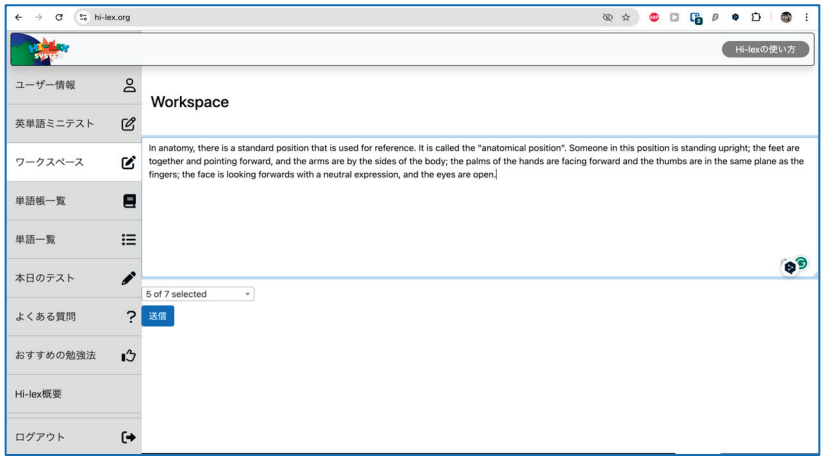

Figure 1. The Hi-Lex browser.

As shown in Figure 2, the software then generates a list of vocabulary along with a simple profile of the words, which the learner can use to decide which words to study. The learner can see whether terms are in key word lists (red circles show inclusion, blue crosses show exclusion, and numbers or letters with numbers indicate inclusion and level). The learner makes the selection by putting check marks in the boxes next to the terms, and can then access the words of their choice to memorise them using spaced-repetition software (Figure 3).

The word lists incorporated in Hi-Lex, which include the TOEIC Service List (Educational Testing Service [ETS], n.d.), JACET 8000 (Japan Association of College English Teachers [JACET], 2003), and CEFR-J (Koda, 2010), are intended for students of general English. To these, the research team added its medical English word list (MEWL). For compatibility with the organisation of Hi-Lex, each item needed both a Japanese translation and an example sentence to provide context for the word. Although there were sentences including most of the word list items in the units, these were not always suitable as a single sentence example, because they were often dependent on the surrounding sentences in the discourse. We decided, therefore, to generate an initial list of sentences using

ChatGPT (https://openai.com/, accessed on 14 February 2025). However, it was necessary to carefully review these sentences to ensure that they fitted with the contextualised meanings of the terms. For example, ChatGPT's sentences for the cryptotechnical terms *superior* and *inferior* had to be re-written to show the anatomical senses of these words. Similarly, a very general-use term such as *mass* needed to be exemplified within the medical context. With fully technical, less context-dependent terms like *astrocytoma* and *urethral sphincter*, constructing sentences was more straightforward. Following a careful check, the items, translations, and example sentences were uploaded into Hi-Lex.

文書中の単語一覧

全選択 全選択解除

登録	英語	文書中の形	日本語	品詞	登録済みの単語帳	cefr	hirotan	jacet	mewl	ngsl
☐	anatomy	anatomy	〈U〉解剖学,解剖論 / 〈U〉〈C〉(動植物の)解剖 / 〈C〉(動植物の)構造,組織;人体	名詞		×	×	8	○	×
☐	standard	standard	〈C〉〈U〉『標準』,基準,水準,規範 / 〈C〉(度量衡の)基本単位 / 〈U〉〈C〉(貨幣制度の)本位 / 〈C〉旗,軍旗;(主義・主張などの)旗じるし / 〈C〉支柱,ランプ台,しょく台 / 『標準の』,標準的な / 定評のある,権威のある	形容詞		B1	○	1	×	○
☐	position	position	〈C〉『位置;場所』;所在地 / 〈U〉所定の位置,適所 / 〈C〉『姿勢』,構え / 〈U〉有利な地位(立場) / 〈C〉《通例単数形で》(周囲の状況と関連する)『立場,境遇』;形勢,情況 / 〈C〉見解,態度 / 〈C〉(…としての)勤め口,職《+『as(of)+名』》 / 〈C〉(…の中の)地位《+『in+名』》 / 〈U〉(特に高い)社会的地位 /	名詞		A2	○	1	×	○

Figure 2. Vocabulary list and word profile generated by Hi-Lex.

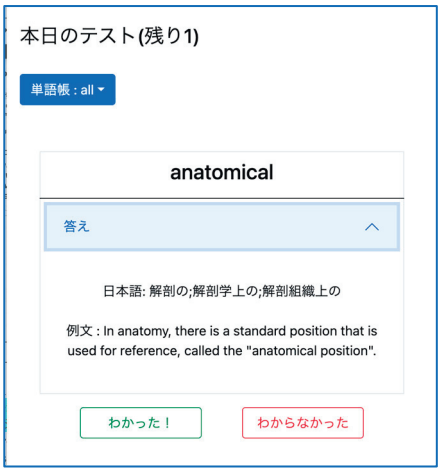

Figure 3. Spaced-repetition software.

7. Trialling Hi-Lex

Initial trials of Hi-Lex in relation to medical English were undertaken from October 2023 to January 2024. The trials covered two seven-week courses, with each course being taught to approximately 35 second-year students.

In the first course, the software was introduced at the start, with 40 students registering on the software. They were given the freedom to use it as they wished. However, the results of the trial showed that only six students used the software, leading to a rethink by the research team. It was decided that the software should be integrated into teaching plans.

In the second course, the instructor used the software by dedicating the last 15 min of each class to reviewing vocabulary. Students were encouraged to copy and paste two of the essays from the online unit PDFs into Hi-Lex for analysis, helping them to identify

the vocabulary items they wanted to learn. This approach to Hi-Lex provided much better results, and it is these that are analysed in this paper (an article presenting a comparative analysis of results from the two courses is currently under review). In addition, students in the second course completed a questionnaire providing feedback on both the course and the software. In total, 28 students registered on the software, and 31 students responded to the questionnaire.

7.1. Students' Saving of Words

A key concern for the research team is the words on which students concentrate, rather than the words selected by the team. Our hope was that there would be a considerable overlap between the two. As an initial trial exploration of students' identification of words for learning, we analysed the saved vocabulary from the 28 students who registered on the software in the second course. Although the software was integrated into the final fifteen minutes of the class, seven students did not use it at all, and three students recorded only four words between them. These ten students are considered to have been uninterested in the software. Among the remaining 18 students, the numbers of words varied considerably, with the average number being 121 with a standard deviation of 115. The minimum number of words for a student was 14 and the maximum was 498. Almost all the words had relevance to the medical English discourse under study during the course, although the student who recorded 498 words had clearly tried a non-medical text, and had recorded words such as *jurisdiction*, *manufacturing*, and *literacy.*

The words recorded by each student (including the three students with only four words between them) were tabulated on a spreadsheet. The words were then collated and sorted according to frequency of recording. A total of 999 words were saved by the students. Of these, 610 items occurred only once, 133 appeared twice, 89 appeared three times, and 61 words occurred four times. For the analysis in this paper, we focus on the remaining 105 items that were recorded more than four times.

7.2. Analysis of Students' Saved Words

In this section, we look in some detail at the characteristics of the items most frequently recorded by the students. We investigate which of these words appear in the MEWL and consider others that are not listed.

7.2.1. What Kind of Words Are in the Students' Lists?

Table 2 lists the words saved by at least five of the 18 students who used Hi-Lex as intended (105 words in total). Words in bold are fully technical terms, italicised words are either lay-technical or cryptotechnical, and those in regular font can be considered to be non-technical, i.e., general/academic, and not specific to medical texts, although they still play an important role. The breakdown of the 105 words by category is as follows:

- Fully technical: 36 words = 34.3%
- Lay-technical/cryptotechnical: 60 words = 57.1%
- Non-technical (general/academic): 9 words = 8.6%

It is interesting to note that while the majority of words listed in the MEWL are fully technical terms, only about one-third of the most frequent words saved by the students fall into this category. More than half of the words are either cryptotechnical or lay-technical—items that we might expect the students to be familiar with. Surprisingly, some non-technical words not listed in the MEWL, including *refer*, *range*, *vertical*, and *consequently*, were selected by students. These are all found in the New General Service List (NGSL, Browne et al., 2013b), which is a list of the most useful and frequently used English words, and they are not words we would anticipate learners at this level to have difficulty

with. It is possible that students chose *refer* because of its potential technical meaning ("refer a patient"), but within the units it was used in a general sense. It was also unexpected that *beneath* was chosen by as many as seven students. When examining the top 30 most frequently occurring words (selected by eight or more students), only six are fully technical, with the most frequent word being a cryptotechnical term (*fuse*). Interestingly, simple anatomical terms like *thumb*, *wrist*, and *elbow* appear on the students' lists. These words were included in the MEWL primarily for completeness, with the expectation that students would not spend much time on them. This finding underscores the importance of carefully selecting words for inclusion in educational materials and emphasises the need to avoid assumptions about which words learners do or do not need.

Table 2. Items with a frequency of over 4: (fully technical [bold font], lay-technical/cryptotechnical [italicised], general/academic [regular font]).

1. *fuse*	(12)	36. *insoluble*	(7)	71. *asthma*	(5)
2. **distal**	(11)	37. **larynx**	(7)	72. **auscultation**	(5)
3. **capillary**	(10)	38. *mold*	(9)	73. *benign*	(5)
4. **fibrous**	(10)	39. **pathogen**	(9)	74. **bile**	(5)
5. **proximal**	(10)	40. **photon**	(9)	75. **cerebral**	(5)
6. reference	(10)	41. refer	(8)	76. *chronic*	(5)
7. *resonance*	(10)	42. *thumb*	(8)	77. *compression*	(5)
8. *anchor*	(9)	43. *tonsil*	(8)	78. consequently	(5)
9. **anterior**	(9)	44. **ultrasound**	(8)	79. *constipation*	(5)
10. *attenuate*	(9)	45. vertical	(8)	80. *dementia*	(5)
11. *bowel*	(9)	46. *vigorous*	(7)	81. *elbow*	(5)
12. **posterior**	(9)	47. *abdominal*	(7)	82. **epithelial**	(5)
13. *ridge*	(9)	48. *agent*	(7)	83. **granuloma**	(5)
14. *rupture*	(9)	49. *canal*	(7)	84. **hemorrhage**	(5)
15. *attachment*	(8)	50. *contraction*	(7)	85. *inferior*	(5)
16. *contagious*	(9)	51. **fistula**	(6)	86. *lace*	(5)
17. *diffuse*	(8)	52. *groove*	(6)	87. **lesion**	(5)
18. *elasticity*	(8)	53. **gyrus**	(6)	88. *membrane*	(5)
19. *expiration*	(8)	54. **hemoptysis**	(6)	89. *navel*	(5)
20. *ingest*	(8)	55. *impede*	(6)	90. **neuralgia**	(5)
21. *inhalation*	(8)	56. *latent*	(6)	91. **osteoarthritis**	(5)
22. *iodine*	(8)	57. *limb*	(6)	92. *palpitation*	(5)
23. *lateral*	(8)	58. **medial**	(6)	93. **paresthesia**	(5)
24. *malignant*	(8)	59. *wrist*	(5)	94. **peritoneum**	(5)
25. *mucus*	(8)	60. *nausea*	(5)	95. **pylorus**	(5)
26. *palm*	(9)	61. **peritonitis**	(5)	96. *radiograph*	(5)
27. **pharynx**	(8)	62. *saliva*	(5)	97. *secrete*	(5)
28. range	(8)	63. **sepsis**	(5)	98. **sulcus**	(5)
29. *residual*	(8)	64. *skull*	(5)	99. *superior*	(5)
30. simultaneous	(8)	65. **systole**	(5)	100. *swallow*	(5)
31. beneath	(7)	66. *tissue*	(5)	101. *tourniquet*	(5)
32. **cortex**	(7)	67. **trachea**	(5)	102. *tuberculosis*	(5)
33. **cranium**	(7)	68. *vessel*	(5)	103. upright	(5)
34. **diastole**	(7)	69. *wheeze*	(5)	104. *vomit*	(5)
35. *fracture*	(7)	70. **arteriole**	(5)	105. within	(5)

7.2.2. Students' Lists and the MEWL

Before comparing the words chosen by the students with those in the Medical English Word List, it is important to note that the current version of the Hi-Lex software operates on the basis of single words. In contrast, the word list itself contains both single-word and multiword terms, meaning that the transfer of the MEWL to the software is not a one-to-one match. For example, the word *resonance* comes from *magnetic resonance imaging*, which is listed as a single item in the MEWL. Also, there were some cases in which Hi-Lex did not recognise words as being part of the MEWL. We are currently checking why *sulcus*, *cerebral*, and *tuberculosis*, for instance, were not identified by the software as MEWL words. However, given these limitations, and that this was a small initial trial of the software, the analysis of the most frequently recorded items yields some interesting insights.

Of the 105 words, 90 (approximately 86%) were covered by the MEWL. Students indicated in their questionnaire responses that they primarily focused on their coursework. Consequently, with the list generated from terms extracted from course materials and single-word MEWL items flagged for reference, most of the words saved by the students were from the MEWL. However, not all the words recorded by the students are found in the MEWL. Two of these words shown in Table 1 are *dementia* and *pylorus*, which are lay-technical and fully technical terms, respectively. *Dementia* is clearly an important term to include in a medical word list, and the students' interest in the word drew our attention to the fact that it had been overlooked. The more interesting term of the two was *pylorus*, which is part of the digestive system. This was neither in the materials nor the word list, meaning that students had added it of their own accord. After confirming with anatomy instructors at the medical school that it was a very important term relating to the stomach, we are revising the anatomical essay and diagram to include the terms *pylorus* and *pyloric sphincter*. In the same way that corpus analysis of medical reference books has revealed important gaps in the course materials, analysing students' word lists and determining patterns can similarly contribute to improving these materials.

The remaining thirteen words, not listed in the MEWL, are not fully technical: *reference*, *anchor*, *ridge*, *range*, *residual*, *beneath*, *refer*, *vertical*, *canal*, *groove*, *consequently*, *lace*, and *upright*. Of these, *canal* (a tubular passage or channel within the body that allows the flow of substances, such as nerves, blood, or other fluids), is clearly cryptotechnical. Two other words, *anchor* and *lace*, are used as verbs in the materials, retaining meanings close to their general uses, though they have cryptotechnical potential when used as nouns. For example, *anchor* refers to a supportive structure that serves as a point of attachment for muscles, ligaments, or tendons, while *lace* refers to an intricate network of fibres or tissues. This illustrates the complexities of how context and parts of speech influence the understanding of word meanings, as well as how these words should be classified.

The other ten words, not classified as cryptotechnical, are more general terms, as discussed in the previous section, and are often used in anatomical descriptions. The cryptotechnical terms should certainly be added to the word list; however, it is harder to determine which, if any, general terms merit inclusion due to their unfamiliarity. For example, *vertical* has a frequency of seven, while *horizontal* has a frequency of four. Although the percentage of such words is relatively small, a larger analysis with more students may help the team to add to the non-technical terms in the MEWL.

8. Reflection

In this section, we consider the success of the project from the perspectives of course delivery, the units of material, the word list, vocabulary study, and the Hi-Lex software. Two of the authors of this paper have been involved in teaching the course, and students

provided feedback via a questionnaire regarding the integration of Hi-Lex, which included questions on the taught units.

8.1. Course and Vocabulary-Rich Materials

The units of materials have been developed from an approach focused on integrating a word list into the content. As we have noted, the process deviated from the initial plan in two important ways: (1) The linear approach, which involved developing materials from a word list created through corpus analysis, was replaced by a parallel process focused on the materials themselves, using corpus analysis techniques to enhance and refine them. (2) Corpus analysis was employed to examine key reference books instead of medical articles. By focusing on body systems, the research team was able to produce a manageable number of units for teaching within the time allocated for an EMP course. This has also allowed teachers to coordinate with each other without duplicating teaching content. While there is a strong focus on vocabulary in the course, this is well integrated into a set of materials designed for use in a communicative approach.

The feedback from students about the course is consistent with that of previous years. Some appreciated the systematic organisation of the material, while others enjoyed developing their communication skills in relation to medicine. One student noted that although it was an English class, there was also an opportunity to gain some basic medical knowledge. One further comment was that the course was well paced, allowing the student to make rapid progress and cover all the material.

8.2. Vocabulary, Word List, and Software

Medical English vocabulary initially emerged through the creation of the materials themselves, which were then enriched through editing informed by corpus analysis. Since most of the words appear primarily in essays and dialogues, they are already contextualised. Key words identified in the discourse are then used in vocabulary exercises in the materials. The construction of the units is helpful for highlighting links between terms and the highlighting of word parts, which are often repeated in different units. *Sub-*, for example, is introduced in the CNS unit in the term *subdural hematoma* and also appears in the musculoskeletal unit with the term *subchondral bone*.

As we have noted, the MEWL and its ongoing development have significantly enriched the materials. This list not only facilitates vocabulary review but also helps to create a word part list that connects to terms embedded in discourse. The introduction of the Hi-Lex system forms the beginning of a move towards making the MEWL much more accessible to students. Although the software cannot currently deal with multiword terms, it provides students with a means for checking vocabulary and saving the single-word items that they want to study. The small trial described in this article illustrates how the software gives insight into the vocabulary students wish to focus on, serving as a tool for greater autonomy in vocabulary learning.

In student comments about the course, several students wrote about how they were able to learn vocabulary, including mentions of collocations and the ways that words can be divided into parts. When asked to rate the helpfulness of Hi-Lex for their learning on a five-point Likert scale, the responses roughly resembled a bell curve: two students rated it as very positive, six as positive, twelve as neutral, eight as negative, and two as very negative. Although this result is slightly disappointing, the trial did expose some weaknesses which we are remedying. In feedback at a conference presentation, a teacher of medical English pointed out the importance of linking the Hi-Lex software to the Moodle system, making it much more accessible to students. In response, for our new courses, we have added a Hi-Lex link in Moodle. Additionally, students indicated that they did not engage

in much English study beyond their medical English classes. Given the pressures faced by medical students, this is not surprising. Analysing texts in Hi-Lex is time-consuming, but it benefits students' learning by allowing them to exercise more autonomy in identifying words from discourse that they wish to learn. For a faster analysis that helps students primarily interested in the MEWL, we are also adding Excel files for the 14 units to Moodle. These consist of the key terms used to create the word list. Students can add them to Hi-Lex and identify terms they want to learn from the files that make up the MEWL.

9. Conclusions

In this paper, we have described the process of creating a pedagogic word list alongside the development of learning materials. The result is a word list containing approximately 1700 terms, integrated into a set of 14 units. We have also considered how the word list is delivered through classroom instruction and two software applications: a Moodle course and our university's Hi-Lex vocabulary learning tool. While the Moodle materials provide a foundational framework for learning, Hi-Lex offers students greater autonomy by allowing them to select items they wish to learn, store these items, and analyse the texts they consider important. Although still in the early stages of development and with some limitations, Hi-Lex has shown potential for allowing learners to customise word lists according to their specific needs, which can be reviewed and used to facilitate the acquisition of specialised vocabulary. The combination of materials, classroom instruction, and software offers learners opportunities to learn medical terms in context and revisit them. To enhance the learner experience with the software, ongoing efforts are focused on improving its functions and usability, including the capability to list multiword terms.

At the outset of our project, our goal was to create materials and a word list to help students make the transition from general English to medical English. With input from medical professors and our own exploration of available resources, we achieved this goal through the interplay of corpus analysis and materials development to create vocabulary-rich content. From this, we developed our medical English word list and a word-part list that links discourse to vocabulary terms and word parts. We have benefited from researching a traditional and very well-established field, but the approach could work for other specialties with a core early curriculum.

An important advantage of Hi-Lex is its ability to offer a clearer understanding of the vocabulary students choose, providing teachers with valuable feedback to help them identify omissions and other challenges in vocabulary learning and teaching. In our case, Hi-Lex may help us pinpoint areas in which the medical curriculum can be strengthened. The software also shows promise in addressing key research questions, such as whether cryptotechnical terms are actually more difficult to acquire than fully technical terms, and whether lay-technical words—or, indeed, general words—pose a greater challenge to ESP learners than previously believed. In addition, we have not yet investigated the use of spaced repetition and other strategies for enhancing word retention and comprehension. Such questions highlight the need for further investigation.

Our future research into the specialised vocabulary of medical English texts aims to address these important issues. At present, the Hi-Lex software is used solely as a research tool at Hiroshima University. However, the concept of software integrating a corpus-analysis tool, linking to established word lists, and incorporating spaced-repetition capabilities, holds significant potential to benefit students in the future.

Author Contributions: Conceptualisation, S.F., M.K.H. and W.D.; Data curation, S.F., M.K.H. and W.D.; Formal analysis, S.F., M.K.H. and W.D.; Funding acquisition, S.F., M.K.H., and W.D.; Investigation, S.F., M.K.H. and W.D.; Methodology, S.F., M.K.H. and W.D.; Software, M.K.H.; Project administration, S.F. and W.D.; Resources, S.F., M.K.H., and W.D.; Supervision, S.F., M.K.H. and

W.D.; Validation, S.F. and W.D.; Visualisation, S.F., M.K.H. and W.D.; Writing–original draft, S.F., M.K.H. and W.D.; review and editing, S.F., M.K.H. and W.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the following Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science: No. 16K02923, No. 19K00823, and No. 20K13155. The authors are very grateful for this support.

Institutional Review Board Statement: Informed consent was obtained from all students, and no personal details were recorded. All data were anonymised completely before analysis. The Hi-Lex activity was integrated into the students' coursework, and had no bearing whatsoever on their grades.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: During the preparation of this study, the authors used ChatGPT (https://openai.com/, accessed on 14 February 2025) for the purposes of generating example sentences to provide context for the words in the Medical English Word List. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Anthony, L. (2018). *Introducing English for specific purposes*. Routledge.

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer software], Waseda University. Available online: https://www.laurenceanthony.net/software (accessed on 14 February 2025).

Belcher, D., & Lukkarila, L. (2011). Identity in the ESP context: Putting the learner front and center in needs analysis. In D. Belcher, A. Johns, & B. Paltridge (Eds.), *New directions in English for specific purposes research* (pp. 73–93). University of Michigan Press.

Bergmann, J., & Sams, A. (2015). *Flipped learning for English Language instruction 4*. International Society for Technology in Education.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (2012). *Corpus-informed research and learning in ESP*. John Benjamins.

Brooks, G. (2023). *Bridging the vocabulary gap for English as an additional language learners*. Hiroshima University. Available online: https://ir.lib.hiroshima-u.ac.jp/00054284 (accessed on 14 February 2025).

Browne, C., & Culligan, B. (2015). *Business service list*. Available online: https://www.newgeneralservicelist.com/business-service-list (accessed on 14 February 2025).

Browne, C., Culligan, B., & Phillips, J. (2013a). *The new academic word list*. Available online: https://www.newgeneralservicelist.com/new-general-service-list-1 (accessed on 14 February 2025).

Browne, C., Culligan, B., & Phillips, J. (2013b). *The new deneral service list: The complete word list and frequency data*. Available online: https://www.newgeneralservicelist.com/new-general-service-list (accessed on 14 February 2025).

Chabner, D.-E. (2023). *Medical terminology: A short course* (9th ed.). Elsevier Health Sciences.

Chan, C. S. C. (2018). Proposing and illustrating a research-informed approach to curriculum development for specific topics in business English. *English for Specific Purposes*, *52*, 27–46. [CrossRef]

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, *15*(2), 102–116.

Cowan, J. (1974). Lexical and syntactic research for the design of reading material. *TESOL Quarterly*, *8*(4), 389–399. [CrossRef]

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

Coxhead, A. (2018). *Vocabulary and English for specific purposes research*. Routledge.

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes*, *51*, 84–97. [CrossRef]

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. [CrossRef]

Davies, W., Fraser, S., Lauer, J., & Howell, P. (2013). English for medical purposes: Teaching an intensive English course to third-year medical students. *Hiroshima Studies in Language and Language Education*, *26*, 49–64.

Educational Testing Service [ETS]. (n.d.). *TOEIC® listening and reading test*. Available online: https://www.ets.org/toeic (accessed on 14 February 2025).

Enokida, K., Fraser, S., Davies, W., & Tatsukawa, K. (2018). Evaluating a flipped learning course for third-year medical students. *Journal of Medical English Education*, *17*(3), 125–129.

Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. CLCS Occasional Paper No. 25. Centre for Language and Communication Studies.

Flowerdew, L. (2015). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching*, *48*(1), 99–116. [CrossRef]

Fraser, S. (2005). The lexical characteristics of specialized texts. In *JALT 2004 conference proceedings* (pp. 318–327). JALT.

Fraser, S. (2006). The nature and role of specialized vocabulary: What do teachers and learners need to know? *Hiroshima Studies in Language and Language Education*, *9*, 63–75. [CrossRef]

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education*, *10*, 127–143. [CrossRef]

Fraser, S. (2009). Breaking down the divisions between general, academic, and technical vocabulary: The establishment of a single, discipline-based word List for ESP learners. *Hiroshima Studies in Language and Language Education*, *12*, 151–167. [CrossRef]

Fraser, S. (2012). Factors affecting the learnability of technical vocabulary: Findings from a specialized corpus. *Hiroshima Studies in Language and Language Education*, *15*, 123–142. [CrossRef]

Fraser, S., & Davies, W. (2022). Medical English curriculum development: The SiReN approach. In H. Glen, J. Falout, & M. Apple (Eds.), *STEM English in Japan: Education, innovation, and motivation* (pp. 161–181). Springer International Publishing.

Fraser, S., Davies, W., & Tatsukawa, K. (2014). A corpus analysis of an anatomy textbook: Preliminary findings and implications for materials development. *Hiroshima Studies in Language and Language Education*, *17*, 119–140.

Fraser, S., Davies, W., & Tatsukawa, K. (2016). Applying internal medicine corpus analysis findings to the development of pedagogical materials. *Hiroshima Studies in Language and Language Education*, *19*, 109–128.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. [CrossRef]

Guest, M. (2022). Prioritizing horizontal spoken discourse in teaching English for medical purposes. In H. Glen, J. Falout, & M. Apple (Eds.), *STEM English in Japan: Education, innovation, and motivation* (pp. 29–48). Springer International Publishing.

Higa, M., & Ashida, Y. (2023). A Pedagogical framework for the development of software for EFL learners. *Hiroshima Studies in Language and Language Education*, *26*, 63–77. [CrossRef]

Hsiao, I. C. V., Hung, S. T. A., & Huang, H. T. D. (2023). The flipped classroom approach in an English for specific purposes (ESP) course: A quasi-experimental study on learners' self-efficacy, study process, and learning performances. *Journal of Research on Technology in Education*, *55*(3), 507–526. [CrossRef]

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454–484. [CrossRef]

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge University Press.

Inman, M. (1978). Lexical analysis of scientific prose. In M. T. Trimble, L. Trimble, & K. Drobnic (Eds.), *English for specific purposes: Science and technology: Education, innovation, and motivation*. Oregon State University.

Japan Association of College English Teachers [JACET]. (2003). *JACET 8000: A new vocabulary list for English education*. JACET.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In *Teaching and learning by doing corpus analysis* (pp. 105–117). Brill.

Koda, K. (2010). *CEFR-J: The Japanese version of the common European framework of reference for languages*. The Japan Society for Educational Technology.

Lam, J. (2001). A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography. In *Research reports* (Vol. 3). Hong Kong University of Science and Technology.

Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). Cambridge University Press.

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

Liu, L. (2022). The contribution of the flipped learning environment to value perception and controllability of classroom activities as antecedents of learners' anxiety: A control-value approach. *Frontiers in Psychology*, *13*, 1000710. [CrossRef] [PubMed]

Lu, Y.-L. (2018). What do nurses say about their English language needs for patient care and their ESP coursework: The case of Taiwanese nurses. *English for Specific Purposes*, *50*(1), 116–129. [CrossRef]

Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, P. (2024). Re-thinking the principles of (vocabulary) learning and their applications. *Languages*, *9*(5), 160. [CrossRef]

Özer, M., & Akbaş, E. (2024). Assembling a justified list of academic words in veterinary medicine: The veterinary medicine academic word list (VMAWL). *English for Specific Purposes*, *74*, 29–43. [CrossRef]

Quero, B., & Coxhead, A. (2018). Using a corpus-based approach to select medical vocabulary for an ESP course: The case for high-frequency vocabulary. In Y. Kırkgöz, & K. Dikilitaş (Eds.), *Key issues in English for specific purposes in higher education* (Vol. 11, pp. 51–75). English Language Education. Springer. [CrossRef]

Schmitt, N. (2010). Formulaic language. In *Researching vocabulary* (pp. 117–148). Palgrave Macmillan. [CrossRef]

Simpson, A. (2022). Medical students' English needs and curricular developments. *Journal of Medical English Education*, *21*(3), 94–100.
Trimble, L. (1985). *English for science and technology*. Cambridge University Press.
Wang, J., Liang, S.-L., & Ge, G.-C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442–458. [CrossRef]
Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, *12*(2), 309–324.
Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, *6*(1), 18–35. [CrossRef]
Wette, R., & Hawken, S. (2016). Measuring gains in an EMP course and the perspectives of language and medical educators as assessors. *English for Specific Purposes*, *42*, 38–49. [CrossRef]
Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27–38. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Languages Editorial Office
E-mail: languages@mdpi.com
www.mdpi.com/journal/languages

Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-4818-8