Effects of print exposure on formulaic language and predictive processing of English

Sean Patrick McCarron

St Anne's College, University of Oxford

Supervised by:
Professors Kate Nation and Victoria Murphy

A thesis submitted to the Department of Experimental Psychology, University of Oxford, in partial fulfilment for the degree of

Doctor of Philosophy
Hilary 2025

Number of pages: 495
Number of words: 54,651 (Chapters), 92,007 (Total)

“Words, English words, are full of echoes, of memories, of associations—naturally. They have been out and about, on people's lips, in their houses, in the streets, in the fields, for so many centuries. And that is one of the chief difficulties in writing them today—that they are so stored with meanings, with memories, that they have contracted so many famous marriages.

[...] They are the wildest, freest, most irresponsible, most unteachable of all things. Of course, you can catch them and sort them and place them in alphabetical order in dictionaries. But words do not live in dictionaries; they live in the mind.”

— Virginia Woolf, 29 April 1937

Front Matter

Dedication

Dedicated to the memories of Professor **Bryor Snefjella** and **Samuel Wiebe,** whom we lost in March 2023 and April 2024 respectively—two extraordinarily intelligent, kind, and thoughtful men whom I admired enormously, and who should still be with us today.

Acknowledgements

Thanks first and foremost to my wonderful family: my dear wife **Lindsay,** and our lovely children **Leif, Iver, Elna** & **Roan**. I love you all dearly. Special thanks to our eldest son, Leif, for painstakingly fixing the formatting problems on the enormous Appendix Table A.5—you are absolutely amazing! A huge thanks as well to my parents, **Sandi** & **Leo**, brother **Peter**, and my in-laws, **Vicki** & **Ken** for supporting a move to the United Kingdom with their four grandchildren and niblings.

Thank you to all my wonderful and brilliant ReadOxford lab members, past and present: **Nicky Dawson, Yaling Hsiao, Matthew Mak, Ellen Taylor, Lena Blott, Mohen Zhang, Emma James, Rainy (Yuzhen) Dong, Jessie (Jinyu) Shi, Tonia Williams, Nicole Law, Catherine (Shuhan) Wang, Alice (Songqi) Li,** and **Rhianna Watt.** I am so thankful and appreciative to have met every one of you. Thank you for your feedback, insight, and encouragement.

I want to thank **Victor Kuperman, Cecilia Heyes, Hamish Chalmers, Gaia Scerif, Kate Watkins,** and **Elizabeth Wonnacott** variously for teaching, guiding, advising, assessing, and/or tolerating me. In particular, I would like to thank both **Nicky Dawson** and **Nick C. Ellis** for agreeing to be my assessors for my Viva Voce, and for providing gracious and insightful critiques of my work.

I also had the privilege of meeting many other lovely people working in the Experimental Psychology department, including (but perhaps not limited to) **Dorothy Bishop, Kathy Sylva, Lucy Bowes, Riddhi Jain, Sam Webb, Malin Karstens, Birtan Demirel, Verena Klar, Sophia Shatek, Bethan Grimes, Fionnuala O'Reilly, Elise Sellars,** and **Athena Chow.** Thanks to all of you for your kindness, guidance, and conversation.

Huge thanks to **Shine (Jeongmyeong) Park** for making it possible to conduct the Korean adaptation of the AFT study (and for the Korean translation of the abstract!), and to **Hui Zhu** for contributing to the French-English translations of the connectives task we used. You were both wonderful to work with. I am also extremely grateful to both **Holly Jenkins** and **Johannes Schulz** for providing guidance on the online eye-tracking implementation, and much of the code used for cleaning and preparing it for analysis.

I also want to thank _____**.** Obviously, I could never forget you—you know what you did!

Finally, I want to express my profound appreciation and admiration for my supervisors, **Kate Nation** and **Victoria Murphy**. It has been the joy and the honour of my lifetime to have shared so many hours in conversation with you both, exchanging words, ideas, and ideas about words.

Declaration and Data Availability

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research. Note that the experimental chapters 2 through 5 maintain the third-person plural "we" to indicate these were in part collaborative efforts, although I wrote the content of these chapters myself.

In line with open science practices, all the materials, data, and scripts associated with this thesis have been made available on the Open Science Frameworks (OSF) website.

Chapter 2

Pre-registration: https://osf.io/8ngwb
Data: https://osf.io/y8926/

Chapter 3

Pre-registration: https://osf.io/nsduz
Data: https://osf.io/q62mt/

Chapter 4

Pre-registration: https://osf.io/zgh8f
Data: https://osf.io/3p6vh/

Chapter 5

Pre-registration: https://osf.io/ec98w
Data: https://osf.io/rftnb

Publications and Copyright Notices

Portions of this thesis have been adapted from the following articles:

McCarron, S. P. (2026). Author Recognition Tests. In H. Nesi & P. Milin (Eds.), *Encyclopedia of Language & Linguistics* (3rd ed). Elsevier.

McCarron, S. P., Murphy, V. A., & Nation, K. (2025). An "Author Fluency Task": Semantic fluency as predictor of L2 vocabulary knowledge. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S136672892510045X

Chapters 2, 4 and 5 will be developed into separate journal articles following the submission of this thesis.

Copyright Notice:

Long Abstract (English)

Written language is distinguished from speech by its more complex and varied vocabulary, grammar, sentence lengths, and more. Consequently, reading experience provides critical exposure to book language, which in turn influences language processing. Because of the relationship between reading experience and language ability, accurately measuring exposure to print (i.e. cumulative reading experience) is often necessary for language research. An Author Recognition Test (ART) is a proxy measure of print exposure in which respondents see a set of names and must indicate only those recognised as published authors. The primary advantage of ART is that it avoids problems of subjectivity and social desirability bias inherent to self-report measures of reading experience, and issues of self-selection and attrition in reading journal studies. Due to the nature of the task, however, new versions of ART must be created and updated regularly for each language and population under study. Additionally, although ART correlates with measures of component skills of reading, including spelling, vocabulary, and reading comprehension for first language (L1) speakers, its reliability and validity are questionable in second language (L2) populations.

This thesis investigates a potential alternative to ART using a semantic fluency task in which respondents name as many authors as possible within a maximum of three minutes. The proposed rationale for this "Author Fluency Task" (AFT) is that speakers often have highly varied sources of language exposure—especially in L2—and these may not be entirely reflected in a closed format test like ART. Moreover, AFT requires an extensive search of explicit memory related to reading experience, whereas ART may be more indicative of general cultural knowledge

which is inaccessible unless prompted. Finally, fluency measures often show evidence of semantic foraging, meaning AFT may allow L2 speakers to cluster names of related authors according to an internal model of categorical information, effectively organising a "mental bookshelf". Taken together, I argue that these features of AFT may be more likely to reflect primary print exposure (i.e. personal reading experience) as opposed to secondary print exposure (i.e. environmental exposure).

In Chapter 2, I report on an initial pilot study of AFT in an L1 English population. AFT was associated with knowledge of book language, as measured by accuracy scores on a lexical decision task featuring corpus-derived keywords of literary fiction, as well as with self-report estimates of reading habits. Comparing the two measures, AFT was a better fit than ART for predicting reading habits scores. Although a model using ART provided a moderately better fit for the lexical decision task compared to AFT, a combined measure using author name selections from both tests ("Author Fluency And Recognition"; AFAR) offered a significant improvement over both ART and AFT, suggesting that although the two measures may both measure print exposure, they likely do so through different means.

In Chapters 3 and 4, I assess AFT's potential for estimating L2 print exposure by measuring its associations with two kinds of formulaic language, namely English discourse connectives and word collocation pairs. These vocabulary items were selected because they are thought to be acquired principally through extensive exposure to written language, although they are common in speech as well. In these chapters, I contextualise the L2 findings with results from a comparison sample of L1 English speakers, both from the general UK population and from Oxford University students.

In Chapter 3, I show evidence that compared to ART, AFT was more highly associated with both connectives and collocations test scores in an L1 French population, whereas the reverse was true in L1 UK English speakers. This dissociation between the two measures suggests ART is finely tuned for native speakers, but may be inadequate for capturing the reading experiences of L2

speakers whose exposure to written English is more varied. Moreover, these relationships between print exposure measures and formulaic vocabulary scores remained significant when accounting for proficiency using the English LexTALE. Conversely, first language print exposure (assessed using French language versions of AFT and ART) and proficiency (LexTALE) minimally contributed to L2 connectives and collocations scores, providing no support for language transfer theories with respect to the acquisition of L2 formulaic vocabulary.

In Chapter 4, I attempt to replicate the findings of Chapter 3 in an L1 Korean sample of English speakers. I found that an L2 AFT and ART were almost equivalently correlated with English connectives and collocations in a sample of L1 Korean speakers, but neither were as strongly associated with vocabulary measures as the L1 French sample in Chapter 3. In this population, L2 English proficiency (LexTALE) was the best predictor overall of English formulaic vocabulary knowledge. Similar to Chapter 3 however, Korean adaptations of both AFT and ART were not associated with L2 connectives and collocations, reiterating a minimal direct role of first language print exposure for learning L2 formulaic language.

Finally, in Chapter 5, I present evidence from an Internet-based eye-tracking study using a novel visual world paradigm task. L1 and L2 participants listened to sentences with either idiomatic or literal English phrases and were shown a set of images for each sentence. On idiomatic trials, target images represented non-decomposable English idioms, and their literal meanings were shown as distractors. On literal trials, this relationship was reversed. Here, I demonstrate that in a population of L1 French speakers, the proportion of anticipatory looks to targets increased as a function of print exposure, suggesting reading experience is a factor in the predictive processing of L2 speech. Furthermore, the same dissociation between language groups found in Chapter 3 was observed here, as compared to ART, AFT was more highly associated with outcome variables in the L2 sample, but the reverse finding was obtained for L1 speakers.

From a practical standpoint, I argue that as a psychometric tool for assessing print exposure, AFT is at worst equivalent to ART in some L2 populations (e.g. L1 Korean speakers), but is likely more informative than ART when assessing language groups which are most typologically and culturally similar to English, or in which learners have more varied sources of English-language exposure (e.g. L1 French speakers). Regarding theory, I argue that these findings are most compatible with emergentist or usage-based theories of language acquisition (Bybee, 2010; Tomasello, 2010; Wulff, 2021).

Short Abstract (English)

Compared to speech, writing offers more complex vocabulary and grammar. Consequently, reading provides essential exposure to "book language" which influences language processing. An Author Recognition Test (ART) is often used to measure first language (L1) print exposure, but may not be as reliable or valid in second language (L2) populations. This thesis explores an alternative method, the Author Fluency Task (AFT), where respondents name as many authors as possible in three minutes. Initial results from a pilot study comparing the two print exposure measures in a sample of L1 speakers showed AFT correlated best with reading habits surveys. Participants were also tested using a novel lexical decision task which used low-prevalence keywords of literary fiction from a large corpus of books. Here, a combined measure of ART and AFT improved models of accuracy scores compared to either measure alone, suggesting the two measures assess print exposure through different facets of memory. Subsequent chapters investigated the use of AFT for measuring L2 English print exposure, focusing on formulaic language. AFT outperformed ART in L1 French / L2 English speakers in models predicting accuracy scores for both discourse connectives and collocations. In comparison, ART was a better predictor for L1 English speakers. An attempted replication in an L1 Korean / L2 English sample showed AFT and ART were virtually identically correlated with both vocabulary scores. In a final study using a novel visual world paradigm, eye-tracking data revealed that print exposure correlated with predictive processing of L2 idioms during speech comprehension. Overall, AFT is a more effective measure of print exposure in certain L2 populations and is equivalent to ART in others. This research emphasises the importance of reading for pleasure for second language acquisition, and provides support for usage-based frameworks.

Short Abstract (French)

Par rapport à la parole, l'écriture emploie un vocabulaire et une grammaire plus complexes. En conséquence, la lecture expose les apprenants au « langage des livres », ce qui influence le traitement linguistique. Les chercheurs utilisent souvent le test de reconnaissance d'auteurs (Author Recognition Test, ART) pour mesurer l'expérience de lecture en langue maternelle (L1), mais ce test peut ne pas refléter de manière fiable l'exposition à l'écrit des apprenants en langue seconde (L2). Cette thèse propose une méthode alternative, le test de fluence des auteurs (Author Fluency Task, AFT), où les participants nomment autant d'auteurs que possible en trois minutes. Une étude pilote auprès de locuteurs natifs montre que l'AFT corrèle avec le « langage des livres » et les habitudes de lecture, tandis qu'une mesure combinant ART et AFT améliore la prédiction dans une tâche de décision lexicale. Les chapitres suivants explorent l'AFT pour évaluer l'exposition à l'écrit en L2, en mettant l'accent sur son lien avec le langage formulé en anglais (connecteurs discursifs et collocations). L'AFT surpasse l'ART chez les locuteurs natifs du français apprenant l'anglais (L1 français / L2 anglais), capturant des sources d'exposition plus variées. En revanche, l'ART prédit mieux les résultats chez les locuteurs natifs de l'anglais. Ces résultats se confirment partiellement dans un échantillon de locuteurs natifs du coréen apprenant l'anglais (L1 coréen / L2 anglais), avec des corrélations similaires pour les scores de vocabulaire. Une étude de suivi oculaire, utilisant un paradigme du monde visuel (visual world paradigm), montre que l'expérience de lecture influence le traitement prédictif des expressions idiomatiques en compréhension orale. Globalement, l'AFT évalue mieux l'exposition à l'écrit pour certaines populations L2, soulignant l'importance de la lecture plaisir pour l'acquisition d'une langue seconde et soutenant la linguistique fondée sur l'usage.

Short Abstract (Korean)

구어체는 문어체에 비해 더 복잡한 어휘와 문법을 사용합니다. 따라서, 독서는 언어 처리에 영향을 미치는 필수적인 "책 언어"에 노출되는 기회를 제공합니다. 독서 노출량을 측정하기 위해 통상적으로 저자 인식 검사(Author Recognition Test, ART)를 사용하지만, 이 검사는 제2언어(L2) 학습자들을 대상으로 할 경우, 독서 노출량에 대한 측정값의 신뢰도가 낮을 가능성이 있습니다. 본 논문은 대안책으로 저자 유창성 검사(Author Fluency Task, AFT)를 탐구합니다. 이 테스트에서는 참가자들이 3분 동안 가능한 많은 저자들의 이름을 말합니다. 모국어 화자(L1)를 대상으로 한 두 가지 독서 노출량 검사들을 비교한 파일럿 연구의 초기 결과에 따르면, AFT가 독서 습관 설문 조사 결과와 가장 높은 상관관계를 보였습니다. 또, 참가자들은 대규모 문학 소설 말뭉치(corpus)에서 저빈도 키워드를 사용한 신규 어휘 결정 과제를 수행했습니다. 여기서, ART와 AFT를 결합한 측정 방식이 단독 검사를 사용했을 때보다 측정 정확도가 더 높아, 두 검사가 기억의 서로 다른 측면을 통해 독서 노출량을 측정한다는 가능성을 시사했습니다. 이후 정형화된 언어(formulaic language)에 초점을 맞추어 L2 영어 독서 노출량 측정 도구로 AFT를 활용하는 연구를 진행했습니다. AFT는 담화 연결어(connectives)와 결합어(collocations)의 정확도 점수를 예측하는 모델에서 L1 프랑스어/L2 영어 화자들에게서 ART보다 더 우수한 결과를 보였습니다. 반면, ART는 L1 영어 화자들을 대상으로 더 높은 예측도를 보였습니다. L1 한국어/L2 영어 화자들을 대상으로 한 복제 연구에서는 AFT와 ART가 어휘 점수와 거의 동일한 상관관계를 보였습니다. 마지막 연구에서는 새로운 시각적 세계 패러다임(visual world paradigm)을 활용해 시선추적 데이터를 분석한 결과, 독서 노출량이 대화 도중 L2 관용구를 이해할 때 예측적 처리와 상관성을 보였습니다. 전반적으로, AFT는 ART에 비해 특정 L2 집단에서 더 효과적인 독서 노출량 검사이며, 다른 집단에서는 ART와 동등한 성능을 보였습니다. 본 연구는 여가 독서가 제2언어 습득에서의 중요성을 강조하며, 사용 기반 프레임워크(usage-based frameworks)에 대한 근거를 제시합니다.

Contents

List of Figures

List of Tables

1. Introduction

Abstract

Written language is distinguished from speech by its more complex and varied vocabulary, grammar, sentence lengths, and more. Consequently, reading experience provides critical exposure to book language, which in turn influences language processing. Because of the relationship between reading experience and language processing, it is important to have an accurate way of measuring exposure to print (i.e. cumulative reading experience). An Author Recognition Test (ART) is a proxy measure of print exposure in which respondents see a set of names and must indicate only those recognised as published authors. The primary advantage of ART is that it avoids problems of subjectivity and social desirability bias inherent to self-report measures of reading experience, and issues of self-selection and attrition in reading journal studies. Due to the nature of the task, however, new versions of ART must be created and updated regularly for each language and population under study. Additionally, although ART correlates with measures of component skills of reading, including spelling, vocabulary, and reading comprehension for first language (L1) speakers, its reliability and validity are questionable in second language (L2) populations. This introductory chapter provides a summary of the research on the relationship between print exposure and language proficiency, explores the psychometric uses and issues surrounding ART, and lays out the rationale for assessing print exposure using a semantic fluency measure for author names, arguing that a recall task of this kind may have certain advantages for L2 speakers in particular.

1.1 Written language and print exposure

Print exposure is an individual's cumulative experience with written language, and serves as a reliable correlate of language skill, including spelling, vocabulary, word recognition, reading comprehension, and many others (Mol & Bus, 2011; Stanovich & West, 1989; Wimmer & Ferguson, 2022; see also section 1.1.1 below). Despite its name, reading experience is not limited to printed, or "physical" books alone. Many of us read books regularly on digital devices such as e-Readers, phones, and tablets, and these activities all fall under the umbrella of "print exposure".[1] In general, researchers are specifically interested in quantifying how often an individual reads for pleasure (sometimes called "free voluntary reading"; e.g., Constantino et al., 1997; H. Kim & Krashen, 1998; McQuillan, 2019; Rodrigo et al., 1996), since it represents additional reading experience outside of work or school.

Granted, most of our daily experiences with written language are likely not encountered on the pages (or screens) of a book. Because the Internet is still largely a text-first medium, it is plausible that online exposure is more highly associated with language skill than exposure to print for most speakers. This possibility has even led to some calls to develop direct measures of social media reading (e.g., Huettig & Pickering, 2019). In fact, however, recent first language (L1) research

[1] Although a recent body of research has suggested some advantages for print compared to digital reading, specifically with regard to reading comprehension (e.g., Baron, 2016, 2020; Jian, 2022; Mangen et al., 2013, 2019; Sage et al., 2020; Singer & Alexander, 2017; in contrast, some report comprehension is equivalent across reading media, e.g., Jeong & Gweon, 2021), this is outside of the scope of this thesis. For now, let us assume that all reading for pleasure is equal. With respect to language proficiency, it is perhaps more important that readers engage frequently with the language of books in their leisure time, regardless of their preferred format.

suggests the opposite—when comparing effects of print exposure, years of postsecondary education, reading attitudes, and website exposure in a study of fluent Norwegian speakers, Strømsø (2023) found that only print exposure predicted reading comprehension scores. Moreover, a high degree of Internet experience negated the positive effects of print exposure for participants with a high degree of both. This suggests that book reading provides a uniquely beneficial kind of exposure, one which is not adequately supplemented through time on the Internet for native speakers.

There are a few possible reasons for a "print" advantage. It is necessary to point out that online text discourse found on social media and discussion forums tends to have more in common with spoken rather than written language, being more conversational and informal (Johns et al., 2020; Snow, 2010). This is important, as there are critical distinctions between writing and speech which have implications for language input. Compared to speech, corpus studies show that "book language" across genres tends to feature greater lexical density and diversity (Berman & Nir, 2010; Roland et al., 2007), as well as longer sentences, and correspondingly, more complex syntax (Biber, 1988), including more passive constructions (Dąbrowska & Street, 2006) and relative clauses (Roland et al., 2007). This increased complexity follows naturally from the disembodied nature of text, which must construct context and meaning *ex nihilo*, whereas spoken language can create meaning through reciprocity and shared context (E. V. Clark, 2020; Snow, 2010). Corpus studies of children's books have also revealed more relative clauses, complex syntax, and greater lexical density and diversity compared to both child-directed speech (Dawson et al., 2021; Hsiao et al., 2022; K. Nation et al., 2022) and

adult television transcripts (Cunningham & Stanovich, 1998). Books are thus not only qualitatively different from other sources of language input, they also distinguish themselves at the earliest stages of reading development.

Yet even within the broad classification of "book language", there is substantial variation. Corpus analyses demonstrate that grammatical complexity varies by genre—for instance, nominalisations (e.g., "the *processing of nominalisations* is more difficult" as opposed to "it is more difficult *to process nominalisations*") feature more commonly in academic writing compared to other varieties (Biber et al., 1998). Additionally, despite evidence that formulaic phrases (e.g., collocations, idioms, lexical bundles, etc.) are more common in speech compared to writing (Biber et al., 2021; Martinez & Schmitt, 2012), certain multi-word expressions occur more often in some forms of academic writing (particularly in relation to university course administration and management; Biber & Barbieri, 2007). The impartiality, formality, precise (or frankly, sometimes ungainly) vocabulary and constructions distinguish academic and scientific writing as its own register, one which can initially seem virtually impenetrable for students (Snow, 2010). With respect to fiction writing, vocabulary use also varies significantly, not only within and between different genres, but also between individual books by the same author (Johns & Jamieson, 2018). These observations have important implications, as the volume and kind of an individual's cumulative vocabulary exposure affects word processing skill. Essentially, the more frequently a word has been encountered, the faster it is recognised (e.g., Balota et al., 2004; Brysbaert & New, 2009; although these effects are generally better explained by the variety of contexts in which a word is found, a measure called "contextual diversity"; Adelman

et al., 2006; and a word's "semantic diversity" improves on this further because it accounts for information redundancy across these contexts; Hoffman et al., 2013; Jones et al., 2012; see also Caldwell-Harris, 2021, for an understanding of how these measures relate). However, word frequency effects also vary in strength according to an individual's personal experience, as evidence shows that higher proficiency readers and those with larger vocabularies are most sensitive to them (Kuperman & Van Dyke, 2013). Like single words, formulaic language like multi-word phrasal units are also sensitive to contextual diversity effects (Senaldi et al., 2022), supporting the view that language may be processed in the same way regardless of the level of the linguistic unit (e.g., Martinez & Schmitt, 2012; Nattinger & DeCarrico, 2010).

To illustrate the relationship between exposure and skill, corpus linguistic evidence has shown that variation in language experience can and should be accounted for in models of human cognition and language, including lexical access and retrieval. Johns et al. (2016) examined the kinds of input that best explain lexical decision response time, using data from the English Lexicon Project (Balota et al., 2007). To conceptualise this idea, imagine two people—one younger, one older—who both select a book to read every day. Both individuals will have their own reading preferences and interests, which may be influenced by demographic traits such as age or culture. Consequently, over time, each person's language exposure will vary according to the books they read. The question is, can we determine which combination of book selections would best explain how fast a young or old reader recognises a word? To answer this, the researchers first amassed several large corpora composed of Wikipedia articles, Amazon product descriptions,

and fiction, non-fiction, and young adult books. Using these datasets, the researchers ran a "hill-climbing" algorithm which randomly sampled portions of each corpus (i.e. sentences of varying length) and fed these "chunks" into a model predicting lexical decision response times. With each pass, the algorithm determined if including a chunk of text improved or harmed the fit of the model. This process continued iteratively until the fit could no longer be improved, at which point it determined the optimal balance of exposure to each dataset. Comparing young and old respondents, Johns and colleagues discovered that the proportions of lexical decision selections from the various corpora differed significantly only between the "young adult" and "fiction" text samples (Figure 1.1). As Johns et al. (2016) write,

> "Given the composition of the different corpora, this suggests that the retrieval time data of these different groups are sensitive to the statistics of different linguistic sources that the subjects have experienced: young adults are better described by simpler examples of language as encoded in young adult books, but older adults are better accounted for by more linguistically diverse fiction and literature books. At least anecdotally, this is consistent with the type of linguistic experiences these subjects likely had." (p. 5)

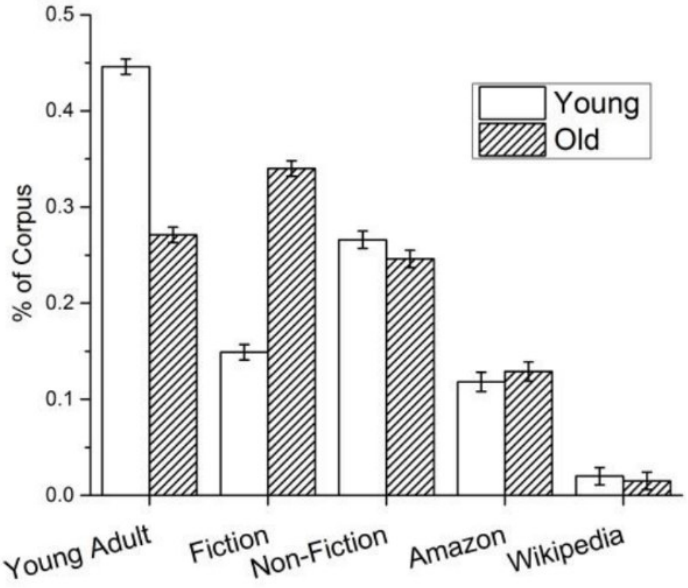


Figure 1.1: Illustration of the differences in selection proportions between younger and older participants. Reproduced with permission from Johns et al. (2016).

To clarify, these findings do not suggest that differences in word recognition between younger and older participants are due to reading experiences alone—rather, they likely reflect individual differences in language exposure more generally, and the divergent vocabulary across book genres simply mirrors the language which is most appropriate for their intended readerships. Yet the finding that varying degrees of exposure to different categories of books can effectively model word recognition times underscores how experience drives language processing.

The importance of exposure aligns with usage-based theories of language acquisition which emphasise that language is acquired through its use, specifically through social interactions and by encountering language in communicative or

meaningful contexts (Ibbotson, 2013; Tomasello, 2000, 2001, 2010; Wulff, 2021), as opposed to the explicit learning of a formal set of rules or definitions. Usage-based theories are situated within the broader cognitive linguistics framework, which posit that language learning relies on domain-general principles of cognition, including embodiment and perception (body and mind are connected, meaning human language reflects embodied interactions and sensory input), categorisation, and memory (Achard & Niemeier, 2004; Lakoff, 1990; Langacker, 1987; Tyler, 2010), rather than appealing to the explanation of a language-specific module in the mind. Although usage-based approaches often focus on the acquisition of syntax (Cheung, 2022), positioning the theory in contrast to nativist notions of a universal, generative grammar (e.g., Chomsky, 2015; White, 1990), they are also relevant for single-word and multi-unit vocabulary learning, both in L1 and L2 (Tyler, 2010). This is because of the robust finding that, "much of language learning is the gradual strengthening of associations between co-occurring elements of the language and that fluent language performance is the exploitation of this probabilistic knowledge" (N. C. Ellis, 2002, p. 173). This notion of language as existing on a continuum of specificity, along which smaller and larger meaningful units (e.g., morphology, syntax, words, and phrases) are processed in a fundamentally similar way according to constraints such as memory, statistical learning, probability, inference, salience, and metaphor, is essential to the perspective of this thesis.

Fundamentally, reading in an alphabetic language such as English relies on the ability to both decode and to comprehend (the "Simple View of Reading"; Gough & Tunmer, 1986; Hoover & Gough, 1990; see also Castles et al., 2018, for an overview of the science of reading). In this framework, decoding is an umbrella

term that encompasses the different processes of meaningfully recognising and interpreting words. When a word is unfamiliar, this decoding involves the learning of phoneme-grapheme correspondences which enable readers to "sound out" words. With time, as reading becomes more fluent, more familiar words are decoded through the faster and more automatic process of word recognition. Naturally, some words will always be more or less familiar to us, and this leads to variations in their "lexical quality"; that is, more or less precise representations of a word's orthographic and phonological form, and its possible meanings in a given context (Perfetti, 2007; Perfetti & Hart, 2002). Greater lexical quality facilitates lexical access and retrieval, leading to more fluent reading, which in turn improves comprehension (Perfetti, 1985, 2007). Yet developing lexical quality is not simply a question of repeated exposure to individual words.[2] To acquire the requisite knowledge about a word's form and meaning(s) to facilitate skilled comprehension, it must be encountered in the company of other words, and in meaningful and varied contexts. Reading provides speakers with precisely these opportunities, and such encounters create a "lexical legacy" of words in the mind (K. Nation, 2017). Words "are full of echoes, of memories, of associations" (Woolf, 1937), and these relations leave behind their own psycholinguistic traces— breadcrumb trails through the worlds readers have visited. Yet these benefits are not limited merely to word recognition. Because of the fundamental importance of language input,

[2] "Words do not often, or even usually, work on their own." (Foster, 2001, p. 91); "Words belong to each other [...]" (Woolf, 1937).

print exposure correlates positively with a panoply of linguistic and non-linguistic outcome measures, a sample of which are summarised below.

1.1.1 Correlates of print exposure

Individual differences in print exposure are associated with vocabulary knowledge (Martin-Chang & Gould, 2008; Mol & Bus, 2011), word recognition (Chateau & Jared, 2000), spelling (Stanovich & West, 1989), collocations knowledge (Dąbrowska, 2014), and synonyms (Mar & Rain, 2015), as well as verbal fluency (Stanovich & Cunningham, 1992), silent reading fluency (Mano & Guerin, 2018), reading comprehension (Martin-Chang & Gould, 2008; Stoops & Montag, 2023), reading frequency (Acheson et al., 2008; Moore & Gordon, 2015), sentence processing (Acheson et al., 2008), and predictive processing of speech (Favier et al., 2021; Huettig & Pickering, 2019). A more comprehensive overview of these effects are described in the meta-analyses by Mol and Bus (2011) and Wimmer and Ferguson (2022).

Exposure to print is also positively associated with academic achievement (Mol & Bus, 2011), fluid intelligence (Fleva et al., 2017), general knowledge (Stanovich & Cunningham, 1993), reasoning (Siddiqui et al., 1998), intuitive comprehension of physics (Black & Barnes, 2015), and the maintenance of executive control with age (Pérez et al., 2022). Additionally, print exposure correlates with a host of purported socio-cognitive benefits including increased empathy, theory of mind (assessed using the "Reading the Mind in the Eyes Test" or RMET, which asks participants to infer mental states from eye expressions, as in Kidd & Castano,

2013, 2017), and social cognition (Fong et al., 2013; Mar et al., 2006; Mumper & Gerrig, 2017), as well as reduced psychological essentialism (i.e. the belief that personal characteristics are fundamentally fixed; Castano et al., 2021) and ethnic prejudice (D. R. Johnson, 2013). However, Samur et al. (2018) notably failed to replicate the Kidd and Castano findings, and Black (2019) raises important psychometric concerns surrounding the use of RMET—originally intended for use in populations with autism—for assessing theory of mind in non-clinical populations. Similarly, Wimmer et al. (2022) did not find effects of reading narrative fiction on social or moral cognition.

Despite the many positive associations between print exposure and other skills, there is continued debate surrounding causality (Martin-Chang & Gould, 2008; Mol & Bus, 2011; Wimmer & Ferguson, 2022). Essentially, reading requires some vocabulary knowledge to begin with, which creates a degree of circularity in reasoning. In other words, do readers have good vocabularies because they read, or do they read because they have good vocabularies? The argument can just as easily be applied to the notion of socio-cognitive benefits to reading—perhaps those with greater empathy are more drawn to literary fiction, which typically features complex character motivations and rich language. Of course, some type of reciprocal relationship is thought to be involved, as it is often claimed that individuals with higher reading proficiency tend to improve their skills faster than less proficient peers. This observation has been dubbed the "Matthew effect" (Stanovich, 1986), in reference to a verse from the book of Matthew in the Bible (25:29) which is sometimes simplified in the popular maxim, *"the rich get richer, and the poor get*

poorer"[3] (but see Protopapas et al., 2016, for psychometric problems with assessing the claim of Matthew effects for reading). The compounding effects of print exposure are argued to make early and sustained reading essential, as its association with oral language skills increases through childhood, adolescence, and young adulthood (Mol & Bus, 2011; Figure 1.2).

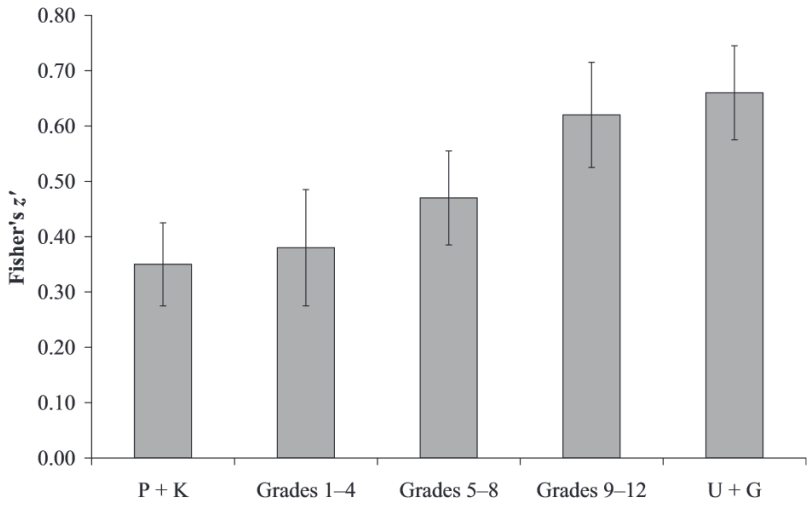


Figure 1.2: Histogram illustrating increasing effect sizes for print exposure on oral language proficiency by education levels. © 2011 by American Psychological Association. Reproduced with permission. Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267–296. https://doi.org/10.1037/a0021890

Complicating this picture, causality models derived from large-scale twin studies show that it is in fact reading proficiency which directly affects print exposure—an

[3] For reference, the actual verse reads, "For to everyone who has will more be given, and he will have an abundance. But from the one who has not, even what he has will be taken away." (ESV Online, 2001, Matt. 25:29)

inversion of the traditional understanding (van Bergen et al., 2018). This implicates genetics as a fundamental catalyst for seeking out reading experiences throughout the lifetime (*"genotype-environment correlation",* e.g. Haworth et al., 2010), leading to calls to control for genetic confounds in home literacy studies (Hart et al., 2021). Yet as van Bergen and colleagues (2018) point out, these findings do not suggest that reading interventions are in vain, or that reading experience does not develop critical language skills. Instead, they remind researchers to avoid the implication of parental blame for inadequate home literacy environments, recognising that there are hereditary factors which may be beyond the control of parents and caregivers. Rather than being rigidly deterministic, then, these genetic findings contextualise reading ability by demonstrating that although it may be influenced through nurture, it can also be constrained by nature.

1.2 Assessment of print exposure

As we have seen, print exposure is positively associated with many language skills, and despite the complex interplay of genetics, personality, and environment, reading experience still plays an important role in shaping language proficiency. Researchers interested in the acquisition of language thus need to be able to accurately assess and measure print exposure across different populations. From a psychometric perspective, objectively quantifying how much a person reads presents certain challenges, however. Due to the perceived value of reading as a pastime, reading habits surveys are considered vulnerable to social desirability bias (Stanovich & West, 1989). Some survey questions may also be interpreted

subjectively, which makes it difficult to meaningfully compare responses from different participants (Mol & Bus, 2011). Other self-report measures have been employed such as reading journals, in which participants must keep detailed records of book reading, yet these are also considered unreliable due to concerns surrounding self-selection and attrition (Mol & Bus, 2011; Sénéchal et al., 1996). To avoid these concerns, researchers assess print exposure using indirect (proxy) measures, of which the most common is the Author Recognition Test (ART; McCarron, 2026).

1.2.1 The Author Recognition Test (ART)

ART is a test that presents a list of names and asks participants to indicate those recognised as published authors (see Figure 1.3). Typically, only some of these names are in fact authors, whereas others are foils (i.e. fake names). To account for possible guessing, the score is most often calculated as the number of author selections minus the number of foil selections.[4] Despite the indirect relationship between author knowledge and reading experience, ART is associated with many component skills of reading, with most of the correlations noted in section 1.1.1 being associated with ART.

[4] Alternatively, this may also be calculated by the percentage of correct (author) selections minus the percentage of (foil) selections (e.g., Brysbaert et al., 2020).

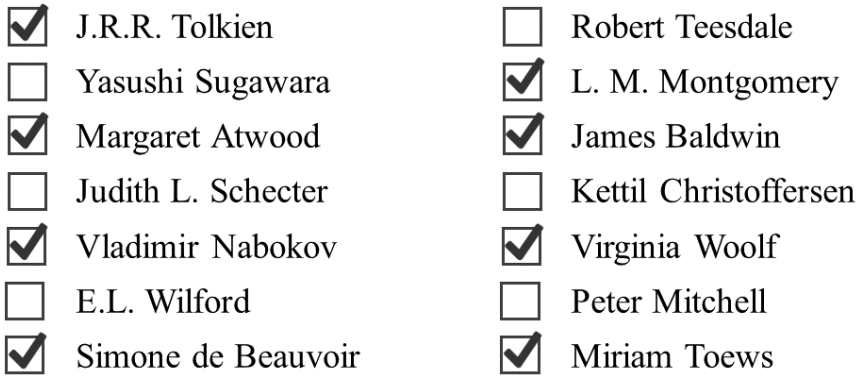


Figure 1.3: An example of an Author Recognition Test, where participants select names of published authors. As a proxy measure of print exposure, ART is associated with component skills of reading.

1.2.2 Versions of ART and related measures

ART was first devised by Stanovich and West (1989), where it was presented in checklist format with 50 author names and 50 foils. Subsequent revisions have been issued in response to changes in author popularity and/or reader familiarity, notably by Stanovich and Cunningham (1992), Acheson et al. (2008), Moore and Gordon (2015), and Vermeiren et al. (2022). Additionally, Mar and Rain (2015) created a “genre” version ("ART-G") that separates scores for fiction and non-fiction author names.

ART has also been adapted for other contexts using author names associated with various languages, including Simplified Chinese (Su et al., 2023), Dutch (Brysbaert et al., 2020), French (Zufferey & Gygax, 2020), German (Hug et al., 2024), Greek (Pérez et al., 2022), Hebrew (Shatil et al., 2000), Korean (H. Lee et

al., 2019), Mexican Spanish (Rodrigo et al., 1996), and Russian (Chernova & Bakhturina, 2023). Additionally, English versions have been adapted for specific regions including Australia (Burt & Fury, 2000), Canada (Chateau & Jared, 2000), and the United Kingdom (Masterson & Hayes, 2007), and a Chinese version for postsecondary students in Taiwan has also been used (S. Chen & Fang, 2015). Author Recognition Tests for children (CART; Cipielewski & Stanovich, 1992; Ricketts, Nation, & Bishop, 2007; Stainthorp, 1997) and for adolescents (Martin-Chang, Kozak, & Rossi, 2020) have also been developed. Similar measures include the Magazine Recognition Test (MRT; Stanovich & West, 1989) and the Title Recognition Test (TRT; Cunningham & Stanovich, 1990), with analogous procedures and scoring methods to ART.

1.2.3 Psychometrics of ART in English-speaking populations

L1 POPULATIONS

There is substantial evidence for ART as an indirect measure of L1 English print exposure (section 1.1.1). Reliability estimates for ART typically range between "acceptable" and "good" (Cronbach's α = .75–.89; Mol & Bus, 2011), and ART demonstrates higher validity than self-report surveys or book counting (Mol & Bus, 2011; Wimmer & Ferguson, 2022). One psychometric criticism of ART, however, is that it may measure more general cultural knowledge rather than reading

experience specifically (Moore & Gordon, 2015; Vermeiren et al., 2022). This may be partly because ART responses do not indicate whether recognising a given author's name represents personal reading *experience* (i.e. an author who the individual has read) as opposed to general reading *exposure* (i.e. an author who the individual has heard of, but not read). This makes it difficult to infer how direct the role of print exposure is—for example, authors such as Stephen King or Jane Austen might be known primarily for the numerous film and television adaptations of their works rather than the original material. Generally, this is not considered to be a problem for ART, as it is understood that additional reading experience increases the likelihood of being familiar with more author names, regardless of whether or not an individual has read a particular author. However, Martin-Chang and Gould (2008) addressed this question of direct vs. indirect reading experience by asking participants to indicate which selected authors they had personally read, and scores for these names ("primary print exposure") were more highly correlated with vocabulary, reading rate, and comprehension than those which were merely recognised ("secondary print exposure").

Evidently then, not all authors on ART are created equal, and it is likely that different authors are associated with different kinds of reading experiences. Accordingly, Moore and Gordon (2015) used item response theory (IRT) to develop a version of ART by determining which names were most discriminative of participants with high and low print exposure, suggesting some names may be a better index of reading experience than others. Moreover, a factor analysis of author names suggested a two-factor structure distinguishing literary and popular authors. Similarly, "ART-G" (see section 1.2.2) has shown that scores for fiction authors are

more strongly associated with verbal language proficiency compared to non-fiction authors (Mar & Rain, 2015). Critically, such differences between fiction and non-fiction reading are not solely the result of differences in word frequency between genres (McCreath et al., 2017), leading some to claim that fiction reading is particularly beneficial for developing language and social cognition because of its emphasis on mentalising and perspective-taking of different characters (elaborated further in section 1.1.1).

L2 POPULATIONS

In comparison to L1, evidence is more limited for ART as an index of second language (L2) print exposure. An IRT analysis by McCarron and Kuperman (2021; building on L1 research from Moore & Gordon, 2015) indicated that ART is unreliable when used in L2 populations and those without a university education. Consistent with claims that tests designed for L1 speakers are less useful in L2 populations (e.g., Vermeiren & Brysbaert, 2023), this study showed that many names on ART are unknown in these groups, creating a floor effect. This is perhaps best illustrated by the item characteristic curves (ICCs) produced from the IRT model, showing responses to items (here, individual author names on ART) by cohort (Figure 1.4). These figures show a line for each author and plot the probability of correctly selecting each one along the y-axis, given the estimated degree of print exposure in each population using standard deviations along the x-axis. Thus, an author name which is likely to be selected at greater than chance probability (represented by the dashed horizontal line) with a relatively lower

degree of print exposure is understood to be a less difficult name, likely reflecting a more well-known author. Conversely, more difficult names, which require a greater degree of print exposure for chance selections, are illustrated as curves that are further to the right on the x-axis. The slope of each curve, in contrast, represents an item's discrimination, that is, how well it distinguishes between a respondent with low or high print exposure, with sharper inclines indicating greater discrimination. In Figure 1.4, these curves illustrate how ART is effective at estimating L1 university students' print exposure at a wide range of ability levels (top left). In contrast, ICCs for L2 university participants (top right) had greater difficulty with author names on ART, as reflected in the rightward shift of curves. Even more seriously, students in a university pre-admission programme for ESL learners (middle left) seemed to find all author names to be of approximately equal difficulty, with the curves mostly aligned further along the x-axis. A reliable and valid test should have items with a range of difficulty values, which in turn can discriminate between respondents of low and high ability, yet ART does not appear to meet this standard in L2.

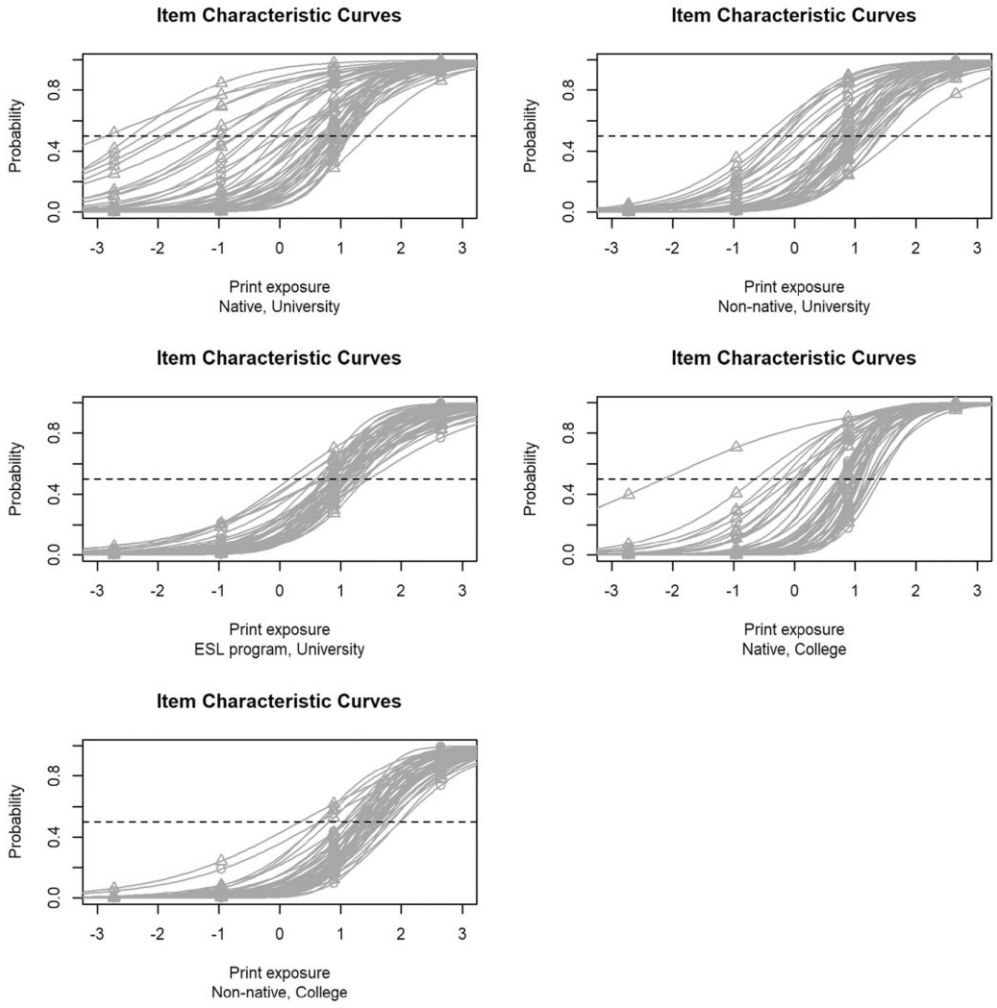


Figure 1.4: Item characteristic curves showing latent trait of print exposure (x-axis) plotted against the probability of correctly selecting each author name (y-axis), for L1 (native) and L2 (non-native) students in college and university programmes in Canada. Each line represents an individual author's name found on ART. (Reprinted from McCarron & Kuperman, 2021).

ART scores also showed decreased measurement precision in L2, particularly at the higher and lower ends of proficiency. As an illustration of this, Figure 1.5 shows a comparison of standard errors of measurement (SEm) for ART across the five

different cohorts. Reliability varies significantly, especially in the lower ranges of ability, and for students without a university education. SEm is also higher for non-native compared to native populations, highlighting a disparity in reliability between groups. Despite its purported unreliability in L2 however, it is worth noting that cross-sectional evidence suggests that the ART scores of L2 postsecondary students improve at a similar rate to L1 speakers during years of post-secondary education, indicating they can close the experience gap with time (a potential "anti-Matthew effect"; McCarron & Kuperman, 2022, p. 58). Nevertheless, longitudinal replication is still needed.

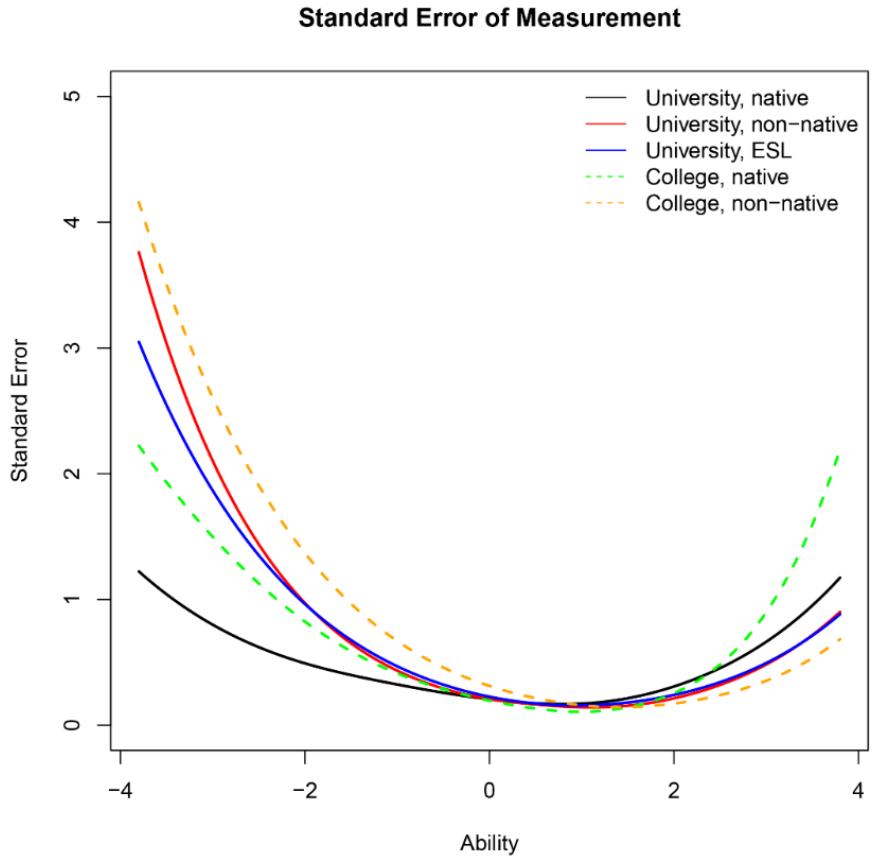


Figure 1.5: A comparison of the standard errors of measurement (SEm) for the Author Recognition Test (ART), across cohorts varying in their levels of education and English language experience (reprinted from McCarron & Kuperman, 2021). Higher SEm is argued to reflect reduced measurement precision.

However, in a response to McCarron and Kuperman, Huang and Bolt (2023) contend that when these data are examined using unipolar (which measure the presence or absence of latent traits using positive scale values only) rather than bipolar IRT models (which measure opposing traits along a continuum with both positive and negative values, centred at zero), measurement error is minimal in lower ranges of proficiency, and greatest in higher ranges. In practical terms, this would mean ART is less reliable when measuring individuals with the most, rather

than the least reading experience. Yet there is reason to remain sceptical about the use of ART as a proxy measure for print exposure in L2 populations. As mentioned, few of the names on ART tend to be recognised by L2 speakers—consequently, high reliability may indicate only that participants tend to recognise the same few names, which may have little to do with primary print exposure, and more with broader cultural knowledge. Because L2 language exposure is considerably more variable and limited compared to L1 (Flege, 2008, 2019; Gullifer & Titone, 2020), it is likely that author names which reflect first language reading experience may not do the same in L2. In other words, ART may in fact be reliable in L2, but not necessarily valid. This is a larger potential issue for ART, since validity is the *sine qua non* of the substantiation of any psychometric instrument—that is, if a test does not measure what it intends to, whether it can do so reliably is effectively meaningless.

In order to improve assessment of print exposure in L2, it is necessary to determine why second language learners tend to not recognise many of the author names on ART. There are likely two types of explanations available—that is, either:

a) L2 speakers generally do not have sufficient reading experience in their second language for author knowledge to be a reliable index of print exposure, or,
b) L2 speakers generally do not read the authors found on ART.

If the answer is a), then ART (or any similar proxy measure for reading experience) is unlikely to work in L2 because it lacks validity. If b) is true, however, then either adapted versions of ART are required for each population being studied, or researchers need a proxy measure which can capture a wider range of reading

experiences. The distinction between these two explanations may prove critical, as the selection of authors who are representative of L1 reading experience may not index the same latent variable in L2. If so, it may be that a valid and reliable measure of reading experience in L2 must acknowledge and exploit this variation, in the parlance of computer programmers, as a "feature, not a bug".

1.3 The Author Fluency Task (AFT)

1.3.1 Rationale

The drawbacks associated with using ART as a cross-linguistic measure of print exposure lead to an obvious question—if the fundamental unit of measurement is the number of author names known by a speaker of a language, why not simply ask participants to list as many authors as possible in a set time? A "verbal fluency" task of this kind would provide a similar proxy for exposure to print as ART, but because it requires more active engagement, it may be more reflective of reading experience.

Verbal fluency tasks generally take two main forms: "phonological fluency" (i.e. listing as many words as possible that begin with a given letter, e.g., *F, A,* or *S),* and "semantic (or "categorical") fluency" (i.e. listing as many exemplars as possible from a given category, e.g., animals, public figures, or items found in a grocery store). In both cases, a participant's score is typically calculated as the sum of unique and valid category items provided in the time allotted. Verbal fluency measures are well-established for assessing semantic and phonological memory, as well as executive functioning (N. Kim et al., 2019; Lezak et al., 2012; Troyer et al.,

1998).[5] Semantic fluency (SF) measures have also been used in L2 studies, generally finding that bilinguals generate fewer category items and proper names than monolinguals. For example, a large-scale study (n = 12,875) of Canadian adults showed that English monolinguals performed better on English animal fluency tasks compared to L1 French bilinguals (Taler & Johns, 2022).[6] In a similar vein, Gollan et al. (2002) reported that Spanish-English bilinguals performed more poorly on a semantic fluency task compared to English monolinguals (both n = 30), whereas bilinguals outperformed monolinguals on a phonological fluency task.

This lower performance on L2 semantic fluency measures undoubtedly relates partially to speed of lexical access, but also to fewer encounters with L2 exemplars; naturally, the two concepts are interrelated, if not inseparable. Accordingly, some evidence shows that highly proficient bilingual adults can in fact perform equivalently to monolinguals on SF tasks (Friesen et al., 2015; Luo et al., 2010). An author naming or production task, then, would rest on the assumption that individuals with greater L2 print exposure can also access more author names extemporaneously, consistent with the "principle of likely need" (Adelman et al., 2006; J. R. Anderson & Milson, 1989; Jones et al., 2017) which relates to how information is retrieved from memory using context.

[5] Because of their associations with executive functioning, semantic fluency tasks are often used in estimating the advancement of neurodegenerative diseases such as Alzheimer's and dementia (Macoir et al., 2006; Troyer et al., 1997, 1998).

[6] Interestingly however, the study also found that L1 English bilinguals performed better compared to monolinguals on the same task, indicating a possible semantic search advantage for bilingualism when testing in L1.

The concept of a semantic fluency task as a potential candidate for an alternative measure of print exposure to ART in L2 has been discussed under different names, initially being described as an "Author Naming Test" (McCarron & Kuperman, 2021, p. 10). Since then, the possibility has been discussed in more recent years (e.g., Hug et al., 2024), and has also been used as an estimate of print exposure in aging populations, using the term "Author Production Test" (APT; Qiu et al., 2024). Yet the concept of an author naming or production task of this kind as an index of reading experience is not entirely novel—in their original ART paper, Stanovich & West (1989) noted that the only predictor of spelling ability on a reading habits questionnaire was whether participants could name just two of their favourite authors.[7] In a sense, the creators of ART had also sketched out the rationale for a spontaneous production task of this kind. To avoid confusion, the term used throughout this thesis is "Author Fluency Task" (AFT), reflecting its nature as a semantic fluency task.

1.3.2 Proposed Advantages

One potential advantage of AFT compared to ART is that it might "level the playing field" for L2 speakers, providing everyone with the same amount of time to demonstrate their print knowledge, whereas not all speakers will have had the same opportunities to encounter the authors on ART. This meets our proposed criteria that a print exposure measure for L2 speakers should capture a broader range of

[7] More recently, this approach was used with children who were asked to provide a few names of their favourite authors, magazines, and books (Spear-Swerling et al., 2010).

reading experiences, and not rely solely on the kinds of authors thought to be associated with L1 reading.

Granted, just like recognising an author on ART, naming an author on AFT may not necessarily reflect personal reading experience. Nevertheless, because SF measures index semantic organisation of memory (Lehtinen et al., 2023), and recognition tasks like ART may reflect "marginal knowledge", or information that is stored but inaccessible unless presented (Berger et al., 1999; Cantor et al., 2015), it is possible that author names which are recalled could be more indicative of primary print exposure than those which are merely recognised. Clearly, author fluency and recognition rely on different skills, with AFT requiring an extensive search of explicit memory related to reading experience, whereas ART is a more passive gauge of recognition. Comparatively, then, AFT might be considerably more challenging for second language learners, since research shows that recall is more difficult than recognition, and particularly in L2 (Vander Beken et al., 2020; Vander Beken & Brysbaert, 2018). Yet this may be advantageous for assessment, since findings have shown that tasks which target productive rather than receptive use of language are more useful for evaluating more advanced learners of English (Webb & Kagimoto, 2009).

Additionally, studies of verbal fluency measures often reveal evidence of "semantic foraging" behaviour, in which respondents "cluster" similar items together until each patch of memory is depleted, then "switch" to another category where additional items are more easily found (Shao et al., 2014; Troyer et al., 1997; Unsworth et al., 2011). In other words, in line with the principle of likely need, the semantic relatedness of items provides a benefit to memory search by allowing

respondents to chain together similar items. This highlights an important difference between fluency and recognition measures of print exposure. With the presentation of each name on ART, participants are forced to begin the memory search effectively from zero, with nothing to link one name sequentially to the next. Rather than "wiping the slate clean", a fluency task such as AFT would allow participants to order and structure their semantic memory in a way that makes sense to them, grouping related author names together into a kind of "mental bookshelf".

Reliability indices also show semantic fluency measures are relatively stable. For instance, Tombaugh et al. (1999) reported that in a sample of older adults who were given an animal naming task twice at an interval of five-and-a-half years (n = 38, time 1 M_{age} = 65.6, time 2 M_{age} = 71.2), test-retest reliability (TRR) was acceptable (r = .74), suggesting scores were mostly consistent in an aging population over a long period of time.[8] Similarly, in a large-scale study of UK citizens (n = 365, M_{age} = 40.75), Harrison et al. (2000) reported a slightly lower TRR (r = .68) for 60-second semantic fluency scores from a subsample of returning participants (n = 90). However, the authors also noted that a longer, 90-second version of the SF task returned a much higher TRR (r = .77), indicating that additional time to respond may improve reliability.[9]

[8] The authors also describe the effects of age and education on their semantic fluency measure, noting that age accounted for far more of the variance on the animal naming task (18.6%) compared to education (13.6%).

[9] This was a consideration in the amount of time provided for AFT, where we ultimately opted to give participants up to 3 minutes (180 seconds) to respond.

1.4 General Discussion

1.4.1 Summary

Exposure to print offers many benefits to language, making its quantification of particular interest to researchers in developmental science and second language acquisition. However, assessing an individual's cumulative reading experience is a complex methodological and psychometric issue. The Author Recognition Test (ART) is a reliable and valid measure for assessing print exposure in L1 speakers of various languages, with evidence that it correlates well with component skills of reading, including measures of spelling, word recognition, vocabulary, verbal fluency, reading comprehension, and more. Additionally, ART has been shown to correlate with skills that are not explicitly related to language, but which are nevertheless important for its acquisition, use, and development, including social cognition and theory of mind. However, questions remain regarding the causal direction of effects, and the application of ART in L2. Moreover, as with any proxy measure, it is difficult to ascertain where any explained variance is derived from; in the case of ART, it is often unclear whether author names are recognised due to personal reading *experience*, or second-hand reading *exposure*. In other words, ART scores may have less to do with print knowledge than with cultural knowledge generally—naturally, it is difficult to disentangle the two.

As a robust associate of many components of language, ART continues to be useful as a quick and easy-to-administer measure of print exposure in many populations, yet there may be ways to improve the assessment of reading experience

in L2 by acknowledging the greater variation in the quantity and kind of second language exposure. One possibility for exploiting this variation is by allowing participants to name authors spontaneously using the framework of a semantic fluency task. In an Author Fluency Task (AFT), participants would be given a set period of time to respond, and scores would be calculated as the number of unique and valid names provided. Since semantic fluency tasks engage explicit episodic memory, it is possible that AFT scores may be more strongly associated with personal reading experience than general cultural knowledge. This possibility will be explored in this thesis by comparing AFT with ART as a predictor of L2 English formulaic vocabulary knowledge.

1.4.2 Thesis Aims

This thesis aims to accomplish the following objectives:

1) Develop a reliable and valid method for assessing L2 print exposure.
2) Determine how print exposure is associated with the knowledge and predictive processing of formulaic language.
3) Describe how this method compares with ART in L1 and L2 populations.
4) Distinguish the roles of L1 and L2 print exposure for acquiring L2 vocabulary.

Each chapter contains a focused literature review, and the chapter outline below serves to orient the reader to what follows:

In Chapter 2, I report on a pilot of AFT in an L1 population by examining the correlations between print exposure and lexical decision scores using corpus derived keywords of literary fiction, as well as with self-report estimates of reading habits.

In Chapters 3 and 4, I assess AFT's potential for estimating L2 print exposure by measuring its associations with two kinds of formulaic language, namely English discourse connectives (e.g. "nevertheless", "furthermore", "given that") and word collocation pairs (e.g. "weak tea", "torrential downpour", "raise prices"). These vocabulary items were selected because they are thought to be acquired principally through extensive exposure to written language. Chapters 3 and 4 are primarily distinguished by the L2 population under scrutiny—Chapter 3 focuses on L1 French speakers of English as an additional language, and Chapter 4 examines L1 Korean speakers of English. In these chapters, the L2 findings are contextualised using results from a comparison sample of L1 English speakers, both from the general UK population and from Oxford University students.

In Chapter 5, I present evidence from an Internet-based eye-tracking study which uses a visual world paradigm task to examine how reading experience in L1 and L2 samples is associated with predictive processing of non-decomposable English idioms.

Finally, the thesis concludes in Chapter 6 with a discussion that reflects on the theoretical and practical implications of this research, and includes some more general and personal reflections in an afterword.

2. Author recognition and fluency predict knowledge of English fiction keywords

Abstract

This chapter introduces the Author Fluency Task (AFT) as a potential alternative to the Author Recognition Test (ART) as a proxy measure of print exposure. In this study, adult, L1 UK English-speaking university students were given 3 minutes to name as many authors as possible, with a point given for each unique and valid name provided, forming the AFT. A standard ART was also administered, which moderately correlated with AFT. Vocabulary was measured with a lexical decision task (LDT) using keywords of literary fiction, and reading experience was assessed using self-report reading habits scores. Results showed AFT and ART were approximately equivalent predictors of reading habits scores, whereas ART was marginally more highly correlated with LDT scores. As an exploratory step, author names from both AFT and ART were also combined into a novel measure, Author Fluency and Recognition (AFAR). A model with AFAR as a predictor of LDT accuracy performed substantially better than either measure alone. These findings are interpreted as evidence that AFT and ART both measure print exposure, but they may measure different aspects of this construct, and through different means. Finally, we propose that as a semantic fluency measure, AFT may have certain advantages over ART—firstly, its open-ended nature may better account for the natural variability in reading experience, since participants are able to indicate authors they have read who may not be included on a standard ART. Secondly, as

semantic fluency measures require explicit memory search, recalled names may be more reflective of personal reading experience (or "primary print exposure") than names which are recognised.

2.1 Introduction

2.1.1 Literary fiction, empathy, and language

As detailed in Chapter 1, print exposure is a measure of a person's cumulative, lifelong reading experience. Because students generally have similar reading opportunities during primary and secondary school years, researchers try to quantify an individual's additional reading for pleasure, that is, reading which is taken on outside of work or education. Fiction writing is thought to contribute to the bulk of this "free voluntary reading", as non-fiction or "informational" texts are commonly read with a particular purpose in mind, such as developing a skill or familiarisation with a given subject. While this may seem unfair to those who enjoy non-fiction reading in their leisure time, research shows that fiction is more highly correlated with verbal proficiency than non-fiction reading (Mar & Rain, 2015), and non-fiction reading habits have even been negatively correlated with reading comprehension in children (Spear-Swerling et al., 2010). This suggests that narrative, as opposed to informational texts, are particularly beneficial in the development of language skill. The reason for this may be partially related to increased familiarity with the contents of fictional stories, which in turn aids comprehension and retention (Frick, 1992; Graesser et al., 1980). As Willingham (2014) writes, stories are "psychologically privileged", making information easier to digest and recall.

There may be other reasons for a possible "story advantage". Among the various fiction subgenres, there has been particular interest in "literary fiction", or simply "literature".[10] Literary fiction tends to focus on the human condition, with multithreaded narratives, advanced vocabulary, and characters with complex and sometimes shifting motivations. The arduous mentalising required to follow literary narratives tacitly invites the reader to fill in gaps left by the author, engaging them in the act of storytelling (Storr, 2021). As touched on in the previous chapter, this may create other advantages, as literary fiction reading is purported to be more strongly associated with many positive cognitive and pro-social characteristics (e.g., Black & Barnes, 2015; Kidd & Castano, 2013, 2017). One of these traits is variously referred to as "mentalising", mind-reading", "social cognition", "empathy", or "theory of mind" in the research literature, although there are arguably subtle distinctions between some of these terms. As the argument goes, the advanced mentalising skills which are honed through reading literary fiction are carried over into the real world.

The link between mentalising and language are well-established, with evidence showing that theory of mind is associated both with L1 language acquisition (Milligan et al., 2007) and L2 language learning in children (Pelletier, 2006). Scores on empathy measures predict advanced L2 speakers' aptitude for producing a "native-like" accent (Guiora et al., 1972; Hu et al., 2013), and advanced proficiency in multiple languages is positively associated with empathy (Dewaele & Wei, 2012). Becoming proficient in a language, whether first, second, or greater, naturally

[10] Or, as Johannes Schulz described it when clarifying the term during a lab presentation, "books they force you to read in secondary school".

requires some degree of mind-reading, as the acquisition of new vocabulary is often a question of inferring a word's referent from context, which may require making inferences about what another person knows or is thinking about when speaking (Bloom, 2000, 2005).[11] Because all three correlate with one another, it is reasonable to suspect that fiction reading, theory of mind, and language skill may be triangulated in some way, [12] even if their causal pathways are less clear. If so, knowledge of the kinds of vocabulary most typical of literary fiction may be particularly useful as an outcome measure when assessing exposure to print.

2.1.2 Individual differences in reading preferences

Individual and group differences in reading preferences may both contribute to outcomes on print exposure measures. Men and women tend to have different reading interests, with reports demonstrating that men are much less likely to read as a pastime in general, and they also read fewer books on average than women (Auxier et al., 2021). Moreover, when they do read, men tend to select books written by men, according to a Nielson Book Research study commissioned for the book *"The Authority Gap: Why women are still taken less seriously than men, and*

[11] Correspondingly, individuals with more advanced forms of autism, a disorder which negatively affects theory of mind (Baron-Cohen, 2000; Tager-Flusberg, 2007), tend to have delayed or disordered language (Bloom, 2000; Eigsti et al., 2011).

[12] Despite this, evidence does not appear to support the conventional belief that autistic individuals prefer to read non-fiction (Chapple et al., 2021; Davidson & Ellis Weismer, 2018).

what we can do about it" (Sieghart, 2021, 2022). As one example, Sieghart writes that just 21% of readers of the celebrated Canadian literary writer Margaret Atwood (author of *"The Handmaid's Tale", "Alias Grace",* and many more) were male. Accordingly, the study revealed a similar pattern for the top 10 bestselling female authors, whereas the readership for top male authors was almost evenly split between genders.

Male readers' androcentric preferences may take root relatively early in life, with recent corpus studies of children's writing showing that although younger children generally prefer to write about characters of their own gender, girls' writing develops a greater balance between male and female characters with age, whereas boys continue to "write about boys" (Hsiao et al., 2021). These differences in creative writing content likely reflect the kinds of stories children personally find most interesting, as well as the content of books they most often read—as Hsiao and colleagues demonstrate, children's books in general still tend to favour stories about boys.

Additionally, men have been shown to read fewer fiction books compared to women (Summers, 2013; Taylor, 2022), suggesting biological and/or social factors may also influence genre preferences. This resistance towards reading female and fiction authors could have negative consequences, not only for these authors and their profit margins, but for both men and women generally, as men may be less likely to understand women's perspectives of the world if they are unwilling to expand their literary horizons (Sieghart, 2021, 2022). Moreover, in addition to being negatively correlated with poorer reading comprehension and verbal ability, as mentioned in the introduction (Mar & Rain, 2015; Spear-Swerling et al., 2010),

non-fiction reading has also been positively correlated with loneliness (Mar et al., 2009).[13] Although these biases in reading preferences may have negative social implications, they may also make it possible to partially validate proxy measures of print exposure by associating author characteristics with demographic attributes of respondents. For example, on an Author Fluency Task (AFT), male respondents might be expected to name more male or non-fiction authors compared to women, reinforcing the notion that knowledge of authors reflects personal reading experiences as opposed to second-hand reading exposure ("primary vs. secondary print exposure"; Martin-Chang & Gould, 2008). To supplement the primary objective of validating a new proxy measure of print exposure, in this chapter we also include an exploratory analysis which tests whether gender differences are evident in the response patterns of participants.

2.1.3 Present Study

As discussed in the previous chapter, the most commonly used measure of print exposure is the Author Recognition Test (ART), but it may be more reflective of general cultural exposure rather than reading experience specifically. Additionally, it has been argued that ART is less reliable when used to assess the print exposure of second language (L2) speakers. In response to these criticisms, the possibility of a semantic fluency task for author names has been proposed as an alternative

[13] This is not to suggest by any means that non-fiction reading is causally implicated in social isolation—more likely, it reflects how existing personality traits manifest themselves through both social interactions and reading interests.

measure of print exposure. The goal of this study was to test and validate the use of an Author Fluency Task (AFT) in a first language (L1) population before evaluating with L2 speakers, which will be explored in subsequent chapters.

In this pilot study, we compare AFT to a standard ART to determine which is more highly correlated with measures of reading habits and vocabulary knowledge. Due to the presumed importance of fiction reading in developing vocabulary skill, we decided to test participants with a novel lexical decision task using keywords of literary fiction, derived from a large corpus of books from various genres (see Methods).

We hypothesised that:

H1: Both AFT and ART would predict reading habits questionnaire scores, but models using AFT would outperform those using ART.

H2: Both AFT and ART would predict lexical decision accuracy, but models using AFT would outperform those using ART.

In our pre-registration, we also planned to evaluate how the print exposure measures predicted response times (RTs) for the lexical decision task, anticipating that individuals with higher print exposure would respond faster to items.[14] However, we discovered that there was virtually no relationship between these

[14] We note that this was not clearly described in our pre-registration, which states that print exposure measures "will positively predict performance (scores and/or response times) on lexical decision task (LDT)". To clarify, this should have stated that print exposure measures would positively predict LDT scores, and negatively predict RTs.

variables (see correlation matrix in Table 2.3.9). We suspect this may have been due to the high level of difficulty for all stimuli, meaning participants had to carefully consider each item. Because of this, we do not explore the relationships between print exposure and LDT response times in greater detail for our analyses below. On an exploratory basis, we also decided to evaluate to what extent author names provided on AFT reflect the demographic characteristics of the individual respondent, testing whether male respondents are more likely to name male and/or non-fiction authors.

2.2 Methods

2.2.1 Participants

This study received ethics clearance from a subcommittee of the University of Oxford Central University Research Ethics Committee [reference R77364/RE001]. Prior to testing, a preliminary power analysis was conducted using the package pwr (Champely et al., 2020) for the R statistical software (R Core Team, 2025). In analysing the relationship between the Author Fluency Task and the lexical decision task specifically, the goal was to obtain 0.8 power to detect a small effect size of .15 at the standard .05 alpha error probability, using a two-tailed test. From these specifications, we obtained a recommended sample size of $n = 346$. For this pilot, we attempted to collect as many participants as possible from a convenience sample at the University of Oxford. We were unable to recruit the targeted sample size, which in hindsight appears to have been an overly conservative estimate. This was likely due to anticipating a much smaller effect size than we ultimately found,

as well as using a two-tailed test. These decisions were informed by the fact that our lexical decision task was a novel measure, so we opted not to make any a priori assumptions about directionality or effects. As detailed below, we were able to find our anticipated effects regardless, and we suspect that the study was not hampered by the relatively smaller sample size.

L1 UK English-speaking participants ($n = 50$, 38 women, 11 male, 1 other, $M_{age} = 19.32$ years) were recruited through the University of Oxford's Experimental Psychology Research Participation Scheme (RPS) between 23 November 2021 and 10 January 2022. Pre-screening required that participants be between ages 18-50 years, native speakers of UK English, and have normal or corrected-to-normal vision. Pilot testing suggested the study would take around 30 minutes per participant. As compensation for completion of the study, participants received course credit.

Prospective participants were first asked to read through the participant information sheet, which provided study details and information about how they would be compensated. Upon consenting, they were invited to follow a link to the Gorilla Experiment Builder website (www.gorilla.sc; Anwyl-Irvine et al., 2020) where they completed the study.

Out of 50 participants, 48 were undergraduate students, 1 indicated they were in graduate studies, and 1 had completed a graduate-level degree. Of the participants who indicated speaking an additional language other than English (22), all reported that they read in English between 76 and 100% of the time. Table 2.2.1 shows a breakdown of self-reported languages spoken by participants. Although all

participants were required to affirm they were native speakers of UK English, the inclusion of some languages which are less-frequently studied in the UK (including Bengali, Dutch, Malay, Tamil, etc.) may raise the question of whether this characterisation is completely accurate.

Table 2.2.1: Additional languages spoken by participants.

Language	**Count**
Chinese	5
German	5
French	4
Spanish	4
Hebrew	2
Korean	2
Polish	2
Russian	2
Other (1 each)	12
Total	38

2.2.2 Measures

PRINT EXPOSURE MEASURES

Author Fluency Task

For the Author Fluency Task, participants were provided up to three minutes to list as many author names as possible, and were given the option to end the task

early if they could not think of any additional names. Instructions asked participants to provide names of authors who had been published in English, though they did not have to be native English writers. Participants were advised the names did not need to be authors who they enjoyed reading, nor those who they had read personally. The instructions also stated that authors were required to be known primarily for their writing—for example, although Barack Obama has written several successful books, they may not have been as well-known were he not also a former U.S. president. The full instructions can be found in Appendix A.1.

Names were typed by participants into a text entry field and each item was recorded after pressing Enter/Return, clearing the field for the next entry. Due to the reportedly difficult nature of the task, names were scored leniently. Each name was checked using an online database of author names (Internet Archive, n.d.) and Google to determine possible cases of unusual misspellings, which were then corrected. Validated author names were rated 1, non-authors (or individuals not primarily known for their writing) -1, and indeterminate names (which could not be verified in the online database or using Google) were rated 0. The coded ratings were then summed for each participant's list of names. For example, a hypothetical participant listing, "J.R.R. Tolkien, Margaret Atwood, Kurt Vonnegut, Conan O'Brien, J. Smith" (respectively 3 authors, 1 non-author, 1 indeterminate), would receive a score of 3-1-0= 2.

Author Recognition Test

This English version of the Author Recognition Test includes 60 fiction author names and 30 foil names ("ART3"; Vermeiren et al., 2022, modified from versions used in Acheson et al., 2008 and Moore & Gordon, 2015). Participants were presented each name serially and in random order, and were asked to respond whether each was an author or not with the keyboard ("F" for yes, "J" for no). Response time and accuracy were both recorded. Correct author selections increased scores by 1 point, and every incorrect selection decreased scores by 1, and no penalty was incurred for not indicating an existing author. The full list of stimuli can be found in Appendix A.6.

Author Fluency and Recognition Measure

Since AFT accepts authors who write in other languages, as well as non-fiction genres, it may be that AFT is better equipped to gauge participants' individual reading experience, while ART measures reading experiences associated with the general English fiction reader. To determine if combining the strengths of both tests might create a more informative measure overall, we combined the author names selected by each participant from both AFT and ART. We then ensured that each name was only counted once, and 1 point was assigned for each name. This exploratory measure is referred to throughout as Author Fluency and Recognition (AFAR).

Reading Habits Questionnaires

The initial version of this measure was a composite score comprised of responses to the following self-report questions about individual reading habits: 1) hours per week spent reading for pleasure; 2) number of books in the participant's home at age 16; 3) number of books read in the past year; and 4) time spent reading the news each day. The maximum score was the sum of the individual measures (28). However, this measure showed poor reliability (Cronbach's $\alpha = .40$). Out of caution, we decided to create an exploratory revised measure of reading habits scores by including additional questions from the reading survey, and excluding items which correlated the least with others (DeVellis, 2017), until we were left with the following items: 1) reading hours per week, 2) books read per year, 3) fiction reading amount, 4) digital vs. print reading percentage. This increased our reliability to the "acceptable" range (Cronbach's $\alpha = .74$). A full list of all questions on both versions of the survey is provided in Appendix A.3.

ADDITIONAL MEASURES

Lexical Decision Task

The list of words used for this task was derived from a large (~240 million words) corpus extracted from published books of various genres (Johns & Jamieson, 2018). To identify words more associated with literary fiction reading, keyness was assessed using the "log ratio" measure devised by Hardie (2014). This measure, also called the *"binary log of the ratio of relative frequencies",* is calculated by taking

the log of the relative frequency in corpus 1, divided by the relative frequency in corpus (equation 1):

$$LogRatio = log\left(\frac{f1}{f2}\right) \quad 1$$

As Hardie explains, the advantage of this method is that it provides a built-in effect size statistic which facilitates the interpretation of results by comparing the magnitude of differences in word occurrence between two corpora. This is in contrast to the log-likelihood method commonly employed in many corpus analyses. Using this method, we indexed words most characteristic of literary fiction, relative to all non-fiction genres.

Literary keywords were selected because of the purported relationship between literary fiction and theory of mind, an important skill for both L1 acquisition (Milligan et al., 2007) and L2 learning (Pelletier, 2006). We then joined these keywords with prevalence norms which estimate how many people are likely to know a given word, indicating a word's difficulty (Brysbaert et al., 2019). Prevalence is also associated with word processing times even beyond effects of frequency, word length, and others (Brysbaert et al., 2019), making it of particular interest. We used these prevalence norms to select lower prevalence (i.e. lesser-known) words, since we reasoned they would likely require greater print exposure for readers to acquire. Finally, we created a corresponding list of non-words using the "Wuggy" pseudoword software (Keuleers & Brysbaert, 2010), which matched our keywords on syllables and number of letters.

As with the ART, LDT stimuli were presented randomly and serially, and participants were asked to respond if they recognised each as an existing English word using the keyboard ("F" for yes, "J" for no). Response time and accuracy were both recorded. Accuracy scores were calculated as the percentage of correct responses to all stimuli (49 words + 47 non-words = 96 total). See Table 2.3.3 for a list of all words used in the LDT, along with prevalence ratings, LogRatio (i.e. keyness), mean accuracy and mean response times. A full list of all stimuli (i.e. including non-words) is also provided in Appendix B.3.

Motivation

The 10-item student opinion survey (Thelk et al., 2009; see also Finney et al., 2016) was used to assess each participant's subjective level of motivation and effort in completing the battery of tasks. The measured variable is the average of all scores, with a maximum of 5. A full list of the stimuli is provided in Appendix A.4.

2.2.3 Procedure

First, participants were presented with information about the study and its aims before providing informed consent. Next, participants completed the demographics and reading habits questionnaires, followed by the Author Fluency Task. Participants then completed either the Lexical Decision Task (LDT) or the Author Recognition Test (ART) before completing the other. For both LDT (word/non-word) and ART (author/non-author), participants saw one item at a time (serial

presentation) and made a yes/no judgement using the keyboard to indicate if they recognised the stimulus as a valid category member or not ("F" for yes, "J" for no). Reaction times for these button presses were recorded, as well as response accuracy. For the LDT, participants were first given practice trials using a similar set of word and non-word stimuli to familiarise them with the task. For the practice phase of ART, participants were shown names of famous actors and non-actors to introduce them to the concept of the name recognition task without using authors. Prior to moving to the trial phase, participants were told that they would next be asked to evaluate whether each name belonged to the category of authors. Finally, upon completion of the main battery of tasks, participants were asked to complete the motivation survey before being granted course credit.

2.3 Results

2.3.1 Descriptive Statistics

In this section, we provide descriptive statistics of each measure, before addressing the research questions related to each in the analysis section which follows. We began by looking at the distributions of our data and computing summary statistics. Figure 2.1 illustrates the density plots for each measure using their original scales.

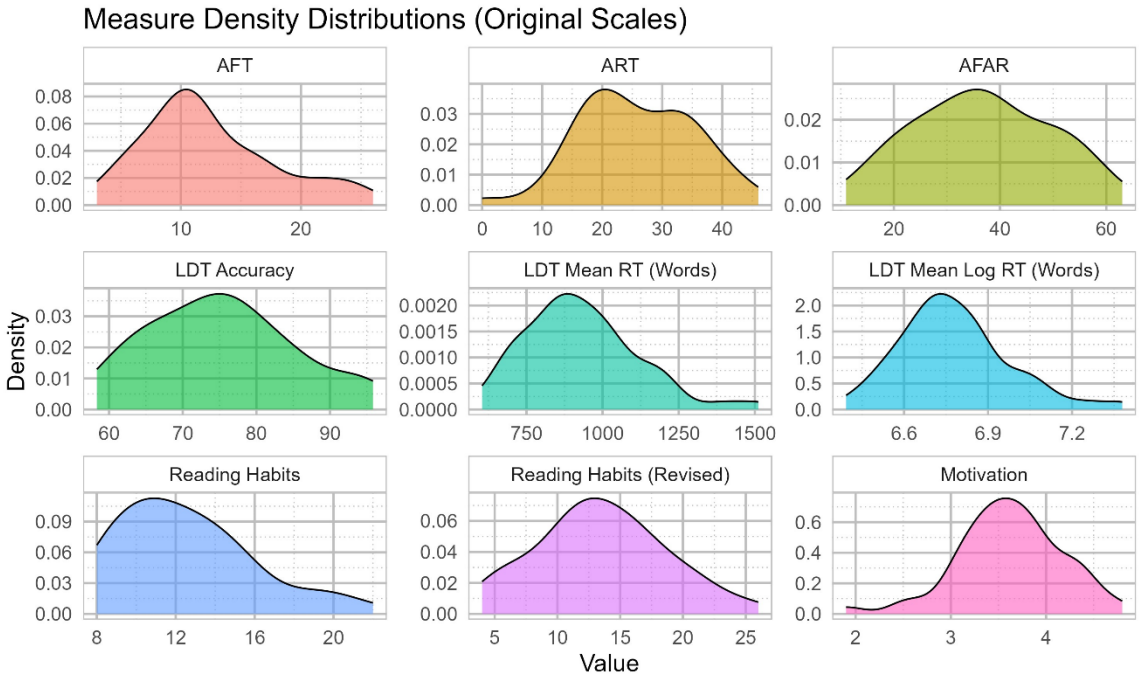


Figure 2.1: Density plots for each measure, in their original scales.

LEXICAL DECISION TASK

Table 2.3.1 shows summary statistics for the lexical decision task in percentages (by lexicality, i.e. word/non-word), and Table 2.3.2 shows the untransformed mean response times in milliseconds to lexical decision stimuli.

Table 2.3.3 shows all words used in the LDT, along with prevalence ratings, LogRatio (i.e. fiction keyness), mean accuracy and mean response times. Response accuracy for words ranged from 26% *(nautilus)* to 94% *(smoky)*, and mean RTs were between 732 ms *(smoky)* and 1253 ms *(incommunicado)*.

Table 2.3.1: Summary statistics comparing accuracy scores (as percentages) on the lexical decision task.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	SE
LDT Accuracy – All Stimuli	58.33	67.97	75.52	75.58	81.25	95.83	10.08	1.43
LDT Accuracy - Words	20.83	45.31	45.31	61.12	76.04	95.83	20.48	2.90
LDT Accuracy - Non-Words	66.67	87.50	91.67	90.04	95.31	97.92	7.07	1.00

Table 2.3.2: Untransformed mean response times in milliseconds for correct responses on lexical decision task, by lexicality (word/non-word).

Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	SE
RTs – All Stimuli	629.40	782.10	874.70	907.80	972.50	1526.90	186.98	26.44
RT - Words	353.00	675.00	805.30	916.00	1027.50	2762.00	407.11	8.22
RT - Non-Words	283.00	645.65	766.00	855.92	955.90	2789.00	383.24	7.91

Table 2.3.3: Keywords of fiction used in the lexical decision task, arranged alphabetically, with prevalence norms (higher = more well-known), LogRatio fiction keyness values (higher = more associated with fiction), response accuracy in percentages, and mean reaction times in milliseconds.

Stimulus	Part of Speech	Prevalence	Prevalence (quantiles)	Log Ratio (keyness)	Response Accuracy (%)	Mean RT (ms)
aspirant	adjective	0.938	3	1.02112	76%	962.51
bawdy	adjective	1.21	5	1.037766	40%	884.09
bereft	adjective	1.116	4	1.06461	60%	831.30
bestial	adjective	1.174	5	1.081971	76%	836.96
clamber	verb	1.036	4	1.073795	86%	867.15
coffer	noun	1.058	4	0.991935	38%	877.72
defrock	verb	1.221	5	1.019224	48%	855.13
denizen	noun	0.689	1	1.03838	36%	771.71
derision	noun	1.027	4	1.03436	56%	867.50
disquiet	noun	1.097	4	1.02841	82%	858.65
epicure	noun	1.060	4	1.032017	72%	976.92
estimable	adjective	0.960	3	1.056108	82%	896.72
exult	verb	0.897	3	1.034015	58%	876.97
feckless	adjective	0.842	2	1.082828	66%	866.85
functionary	noun	1.196	5	1.05709	90%	888.42
genteel	adjective	0.839	2	0.99994	34%	901.47
gristly	adjective	0.635	1	1.024446	72%	1007.29
gumshoe	noun	0.795	2	1.073701	40%	974.79
hackneyed	adjective	0.847	2	1.087996	38%	1046.24
halitosis	noun	1.19	5	1.043688	40%	928.24
hangar	noun	1.167	5	1.083413	44%	922.27
hellcat	noun	1.053	4	1.036781	44%	912.23
heraldic	adjective	0.829	2	1.002165	70%	844.43
hothouse	noun	1.194	5	1.099792	58%	977.92
ignominiously	adverb	0.844	2	1.085382	44%	1162.43
imbecility	noun	0.945	3	1.045219	78%	1030.63
incommunicado	adjective	0.939	3	1.087454	28%	1252.95
lackey	noun	1.216	5	1.072472	54%	907.99
lichen	noun	1.173	5	1.057287	68%	880.49

Stimulus	Part of Speech	Prevalence	Prevalence (quantiles)	Log Ratio (keyness)	Response Accuracy (%)	Mean RT (ms)
longshoreman	noun	1.141	5	0.995806	48%	1150.08
luridly	adverb	0.953	3	1.006644	58%	992.44
madcap	adjective	0.801	2	1.013746	40%	979.17
malady	noun	1.084	4	1.056113	54%	929.24
misspent	verb	1.074	4	1.074966	78%	890.67
nautilus	noun	1.060	4	1.019418	26%	882.38
necromancer	noun	0.970	3	1.01367	76%	891.29
ornery	adjective	0.927	3	1.084818	30%	885.48
ostentation	noun	1.170	5	1.025032	82%	964.94
pique	noun / verb	1.163	5	1.030216	74%	857.96
plebeian	adjective / noun	1.030	4	1.05056	62%	884.97
prepossess	verb	0.868	2	1.089418	72%	1185.80
resplendent	adjective	1.007	3	1.094988	48%	996.22
rummy	noun	1.159	5	1.060364	56%	884.12
smoky	adjective	1.203	5	1.035985	94%	731.83
snugly	adverb	1.205	5	0.991433	62%	893.24
thrall	noun	0.996	3	1.033626	68%	827.75
totter	verb	0.896	3	1.039218	62%	946.21
unexceptionable	adjective	0.770	1	1.081049	90%	1040.26
venturesome	adjective	1.196	5	1.066198	76%	1030.17

PRINT EXPOSURE AND MOTIVATION

Table 2.3.4 provides summary statistics for raw scores of each print exposure measure described below, including the "AFAR" measure which combines unique author names produced in the AFT and recognised in the ART into a single index.

Because it is used as a covariate in some of our best-fitting models, we also include summary statistics for the motivation task.

Table 2.3.4: Summary statistics comparing raw scores on print exposure and motivation tasks.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	SE
Motivation	1.90	3.30	3.65	3.62	3.98	4.80	0.55	0.08
Reading Habits	8.00	10.00	12	12.62	15.00	22.00	3.45	0.49
Reading Habits (Revised)	4.00	10.25	13	13.54	16.75	26.00	5.08	0.72
ART	0.00	19.00	25	25.84	32.75	46.00	9.49	1.34
AFT	3.00	9.00	11.50	12.36	15.00	26.00	5.49	0.78
AFAR	11.00	27.25	36.50	36.70	46.50	63.00	12.80	1.81

Author Fluency Task

First, names provided on AFT were validated and scored. To do this, all names were changed to lower case, removing punctuation and excess spaces to create consistent formatting. Next, clear misspellings were corrected, e.g., "jr tolkein" would become "jrr tolkien". Immediately recognisable author names were assigned a score of 1, leaving a list of unverified names which were checked using the author

databases on OpenLibrary.org and Goodreads.com (*Goodreads*, 2022; Internet Archive, n.d.). If names were not found on these sources, they were searched on Google with the keyword "author" to look for evidence of an author with a similar name. A lenient scoring procedure was used, such that any reasonable approximation of an existing author's name was accepted. Names which did not belong to a published author, and/or which instead belonged to a public figure known for something other than writing were assigned a score of -1; indeterminate names were assigned a score of 0. Duplicate author names from the same participant were also removed. Table 2.3.5 lists the ten most frequently named authors on AFT, along with the number of times each was provided.

Table 2.3.5: Ten most frequently named authors on the Author Fluency Task.

Order (by count)	**Author Name**	**Count**	**Percentage**
1	J.K. Rowling	40	80%
2	William Shakespeare	28	56%
3	J.R.R. Tolkien	19	38%
4	Charles Dickens	18	36%
5	Roald Dahl	17	34%
6	Jane Austen	16	32%
7	Jacqueline Wilson	15	30%
8	Enid Blyton	11	22%
9	Charlotte Brontë	9	18%
10	C.S. Lewis	9	18%

An important unanswered question for validating the use of AFT as a measure of print exposure is whether or not the author names provided are likely to reflect personal reading experience. To partially address this question, we first coded each

of the authors listed by type (fiction/non-fiction), the author's gender (male/female), and the genre most associated with the author's published works. Given extensive reports of gender biases in reading preferences (Auxier et al., 2021; Sieghart, 2021; Summers, 2013; see section 2.1.2 for more), and research evidence of androcentrism in corpus studies of children's writing (Hsiao et al., 2021), we asked whether a participant's gender was associated with the proportion of male or female author names provided. First, a Wilcoxon rank sum test confirmed that males and females did not significantly differ in their overall AFT scores ($W = 192.50$, $p = .70$). We then evaluated the ratios of male-to-female author names provided by male ($M = 4.89$, $SD = 3.94$) and female ($M = 1.45$, $SD = 1.06$) participants. Next, we created a 2 x 2 contingency table of participant and author genders (see Table 2.3.6). We then conducted a chi-square test of independence to examine the relationship between participant gender and author gender counts, revealing a significant association, $\chi^2(1, n = 614) = 26.01$, $p < .001$, indicating that male participants were more likely to name male authors.

Table 2.3.6: Contingency table of participant and genders of authors named on AFT.

Participant Gender	**Author Gender**	
	Female	*Male*
Female	212	257
Male	33	120

Similarly, in light of reports that women may read more fiction than non-fiction books compared to men (Summers, 2013; Taylor, 2022), we examined whether these

preferences might be reflected in the author names they provided. We first evaluated the ratios of fiction-to-non-fiction author names provided by male (M = 11.59, SD = 11.11) and female (M = 13.80, SD = 10.36) participants, which suggested a slight fiction preference among women. We then created a 2 x 2 contingency table for participants by gender (male/female) and the type of book most commonly associated with each author's name (fiction/non-fiction), see Table 2.3.7. A chi-square test of independence showed a significant relationship, $\chi^2(1, n = 614) = 14.53$, $p < .001$, suggesting that male and female participants did vary in terms of the number of non-fiction and fiction authors named. Specifically, although they were somewhat underrepresented in this sample, males named a relatively higher proportion of non-fiction compared to fiction authors.

Table 2.3.7: Contingency table of participant and author genres named on AFT.

Participant Gender	**Reading Type**	
	Fiction	*Non-Fiction*
Female	427	36
Male	122	29

Finally, we asked whether a participant's gender was also associated with genre preferences more broadly. We created a 2 x 24 contingency table for participant gender by the genre most associated with each author (e.g. literary, science fiction, romance, history, philosophy, etc.). Given the small cell values for many of the genres, we conducted a two-sided Fisher's exact test, which showed a significant

difference between the two groups ($p < .001$), showing men and women tended to name authors from different genres.

With just 11 male participants, it is unclear how generalisable these findings are. Nevertheless, they appear to reinforce previous findings that gender and reading preferences are associated, warranting further investigation. More importantly for the present discussion, however, these analyses provide some preliminary indication that the names provided on the Author Fluency Task are likely to reflect personal reading experiences, or at least the kinds of author knowledge we would expect to be more associated with male or female readers.

AFT also showed moderate Spearman correlations with ART ($\rho = .40$), LDT accuracy ($\rho = .39$), reading habits ($\rho = .43$) and revised reading habits ($\rho = .54$, see Table 2.3.9), providing support for the convergent validity of the measure. Because these correlations are modest, it may also suggest that the different measures are tapping different aspects of the same underlying association. Moreover, AFT shows good divergent validity, as it does not correlate strongly with unrelated measures such as the motivation score ($\rho = .17$), whereas ART does ($\rho = .44$). This is important, because convergent validity can only be determined if a measure correlates more strongly with theoretically associated measures than those which are unrelated (Wimmer & Ferguson, 2022). This is arguably even more true for proxy measures, which by their nature absorb variance from numerous indirect factors, making it difficult to determine which latent variables they represent.

Ideally, we would have also calculated statistics for reliability and measurement precision for AFT—however, due to the nature of semantic fluency tasks, this would require re-recruiting participants to take AFT again in order to determine the task's test-retest reliability, which was not feasible for this study. However, in a subsequent study, detailed in Chapter 5, we report that in a sample of the general UK L1 English population ($n = 50$) the measure has good test-retest reliability ($r = 0.80$, SEm $= 2.55$). Given that the population of Oxford undergraduate students in the present study is likely to be more homogenous than the general UK population (an observation often referred to as the "range restriction" problem, e.g., Sackett & Yang, 2000; Vermeiren & Brysbaert, 2023), we expect that reliability may have been somewhat higher here, although this is simply conjecture.

Author Recognition Test

The full list of authors on ART, along with mean reaction times and selection percentages is provided in Appendix B.1.1. ART scores were normally distributed (Shapiro-Wilk test, $p = .67$). Only 24 of 60 author names (40%) were selected more than 50% of the time, despite authors comprising the majority of stimuli (66.67%), suggesting relatively few of the names on ART are familiar to these participants. This is not inherently a problem, as a well-constructed test requires both easy and difficult items to discriminate between respondents of high and low ability (Moore & Gordon, 2015); rather, we note this merely for descriptive purposes. Figure 2.2 illustrates the number of correct author selections on the task.

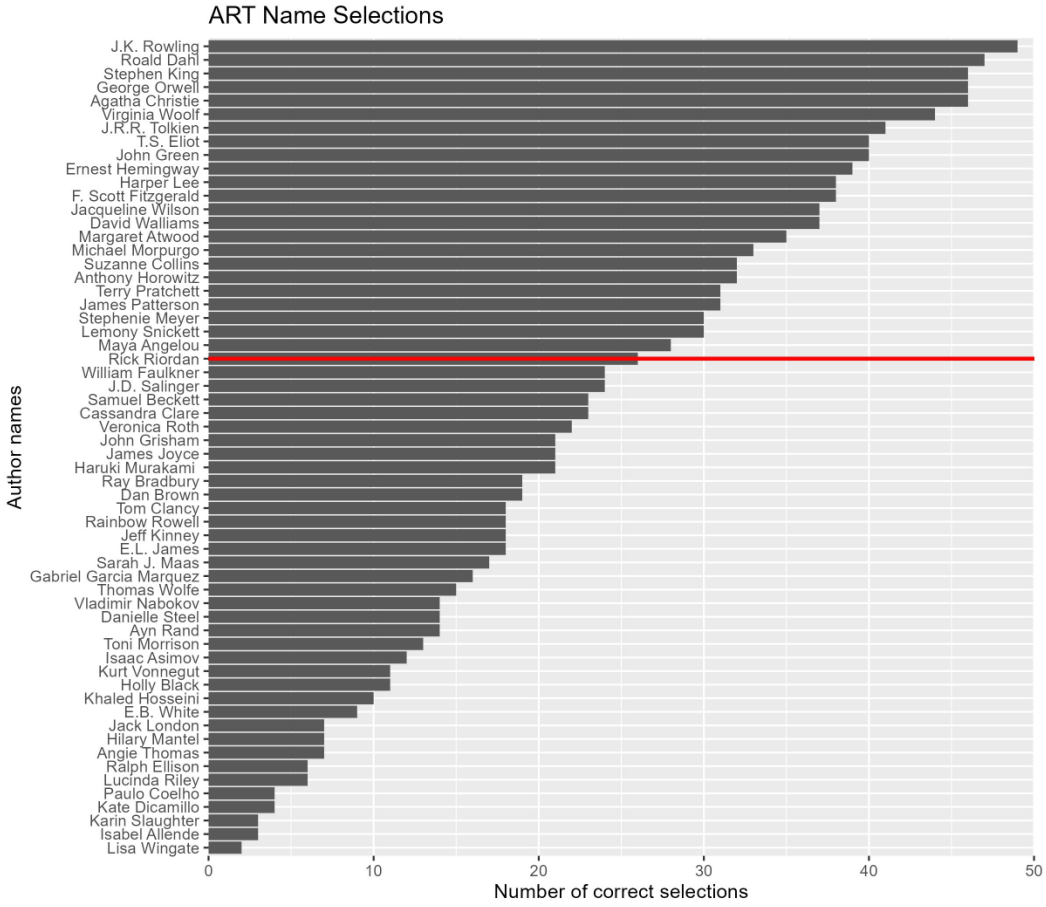


Figure 2.2: Number of author names ($n = 60$) selected on ART by all participants ($n = 50$). Names at or above the red line (Rick Riordan) are those with a recognition rate greater than 50%.

After evaluating the distribution of reaction time observations, an exploratory polynomial regression analysis was conducted to describe the relationship between the percentage of times each author was correctly selected and corresponding mean log-transformed reaction times for each name, across all trials (i.e. RTs for both correct and incorrect selections of each author name). This model was significant, ($F(2,\ 57) = 21.12,\ p < .001$; see output in Table 2.3.8), and accounted for

approximately 45% of the variance in mean reaction times ($Adj\text{-}R^2 = .45$). Exponentiating the coefficients to interpret them on the original scale, the analysis revealed a significant positive linear effect of author selection accuracy ($\beta_1 = 1.01$, $t(57) = 4.19$, $p < .001$) and a significant quadratic effect ($\beta_2 = 1.00$, $t(57) = -5.45$, $p < .001$), indicating an inverted U-shaped relationship between accuracy and reaction times (as illustrated in Figure 2.3). Essentially, this curve describes how author names which are least and most frequently recognised are also responded to the fastest, with names in the middle range taking significantly longer. This could demonstrate one advantage of conducting an ART more like a lexical decision task as opposed to the traditional checklist format, as it enables researchers to evaluate processing speed for each name individually.

Table 2.3.8: Model output from a polynomial regression analysis predicting mean response times to author names by author name selection accuracy. For interpretability, we have exponentiated the original output for our coefficient estimates and confidence intervals.

	Log Mean RTs ~ Author Selection Accuracy		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	801.03	747.685 – 855.943	$<$ **0.001*** **
Author Selection Accuracy	1.006	1.003 – 1.008	$<$ **0.001*** **
Quadratic Term	0.999	0.999 – 0.999	$<$ **0.001*** **
Observations	60		
R^2 / R^2 adjusted	0.465 / 0.446		

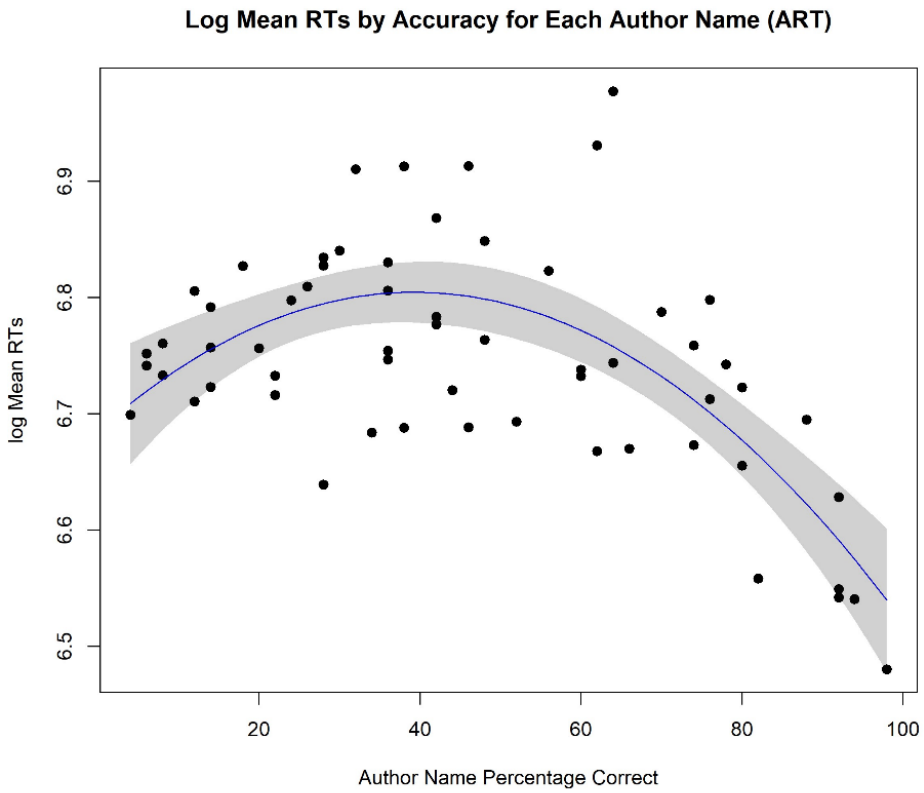


Figure 2.3: Scatter plot showing the relationship between the percentage of correct responses (Author Name Percentage Correct) and log mean reaction times for each author name (for all trials, i.e. averaged RTs for both correct and incorrect responses), represented by black dots. The curved line represents the quadratic fit of the data, and the shaded area indicates the 95% confidence interval of the fit.

We calculated several additional estimates to assess reliability and precision of measurement on the overall ART, and with respect to authors and foils alone. The internal consistency of the test was evaluated using the Kuder-Richardson Formula 20 (KR-20), yielding a reliability coefficient of .91 (authors = .92; foils = .76), which indicates a high level of consistency in participants' responses across items, in particular for recognising authors as opposed to foils. The Standard Error of

Measurement (SEm) was also calculated to assess the precision of individual scores. The SEm for all items was 2.76 (authors = 3.04, foils = 1.26), meaning that the observed scores were expected to fluctuate by approximately 2.76 points due to measurement error, although the greater precision for foils could be interpreted as reflecting minimal guessing for these items. Additionally, the Intraclass Correlation Coefficient (ICC) was computed to evaluate the consistency of scores across items. The ICC2k was .78 (authors = .88; foils = .74), indicating substantial agreement and reliability of the test scores. Collectively, these estimates demonstrate that this version of ART is a reliable measure of print exposure in this L1 population, consistent with robust findings that ART performs well in native English-speaking, university-educated populations (McCarron & Kuperman, 2021).

Comparing AFT and ART

We also evaluated the degree to which names provided on AFT varied from those recognised on ART. Participants as a group named a total combined list of 261 valid and unique authors on AFT, and all 60 authors were selected at least once on ART. Combined across both measures, a total of 283 valid and unique author names were selected. Of these names, only 22 were found on ART but not AFT (7.77%), whereas 223 were found on AFT but not ART (78.80%), visualised by the Venn diagram at right in Figure 2.4. As illustrated in the left of the same figure, of the 261 names provided on AFT, the majority (175, or 67.05%) were named only once across all participants. This may reflect a significant degree of individual variation in participants' reading experience, even in this relatively homogenous

cohort. Of course, an open-ended fluency task like AFT will naturally provide a greater quantity of items overall, so this observation alone is not surprising. However, we suggest that it reinforces the notion that names provided on AFT are likely more representative of individual reading experience than those selected from ART.

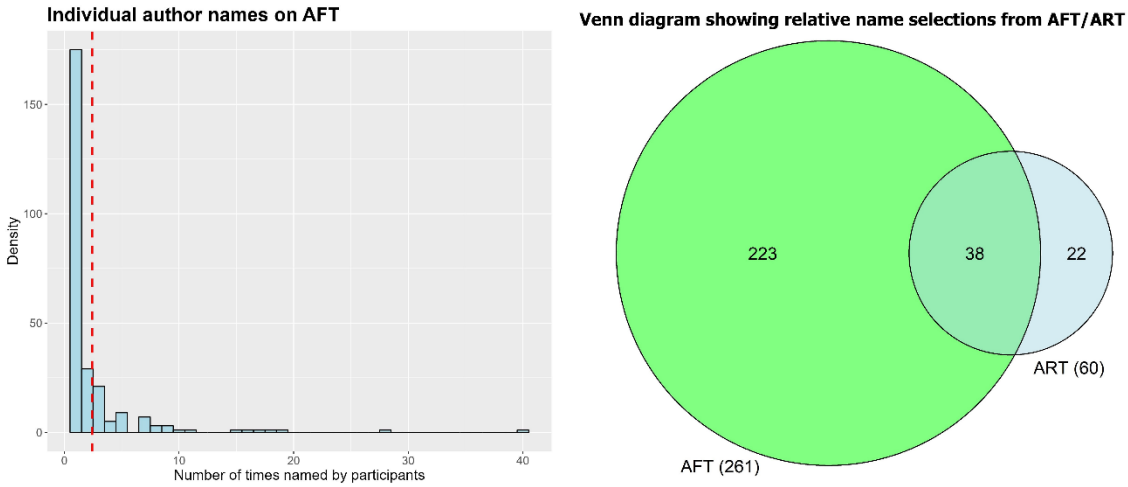


Figure 2.4: (Left:) Histogram illustrating the number of times each individual author name was provided on AFT, showing a Zipf-like distribution. The red dashed line indicates the mean. (Right:) Venn diagram of unique author names found on AFT and ART.

Reading Habits

Bar plots of individual questionnaire items are shown in Figure 2.5. Question 1 showed that most participants read for pleasure between less than an hour and three hours each week. Question 2 revealed that almost all participants had at least 26 books in their homes at age 16, with many respondents indicating between 100-

500. Participants also indicated they had read between 1 and 10 books in the past year, consistent with market research about the general UK population (McClelland & Powell, 2021). Finally, Question 4 showed that most participants read less than 30 minutes of news each day, with most reporting they did so for 1-15 minutes.

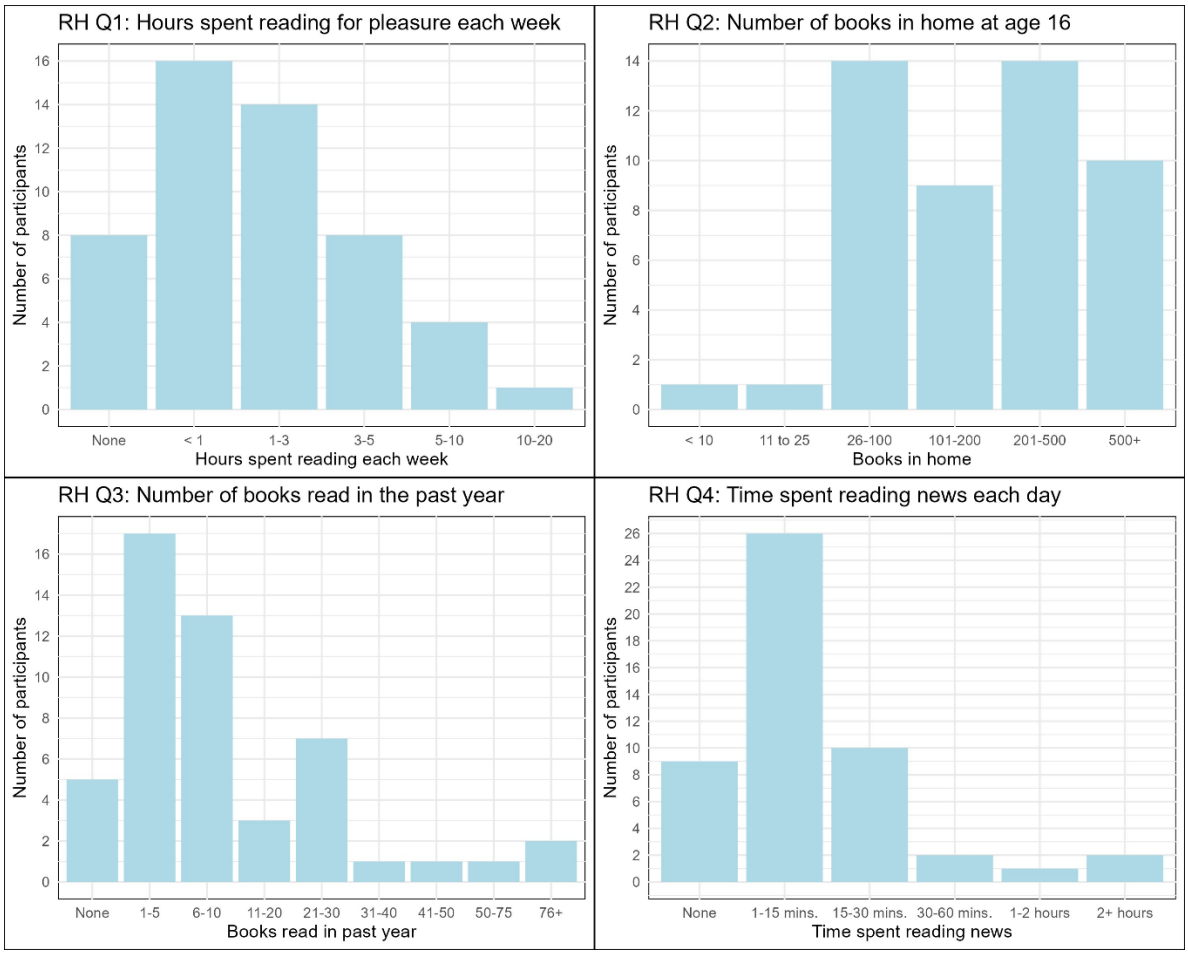


Figure 2.5: Bar plots showing responses to self-report reading habits questions forming the initial version of the reading habits questionnaire.

Additional questions added to the revised questionnaire are shown in Figure 2.6, with Question 5 ("Percentage of fiction vs. non-fiction reading") and Question 6

("Percentage of digital vs. print reading") replacing Question 2 ("Books in home") and Question 4 ("Time spent reading news"). Responses showed that the majority of participants tended to read fiction during their leisure time. Additionally, Question 6 revealed a moderate advantage for print (62% of respondents) over digital (38%) reading.

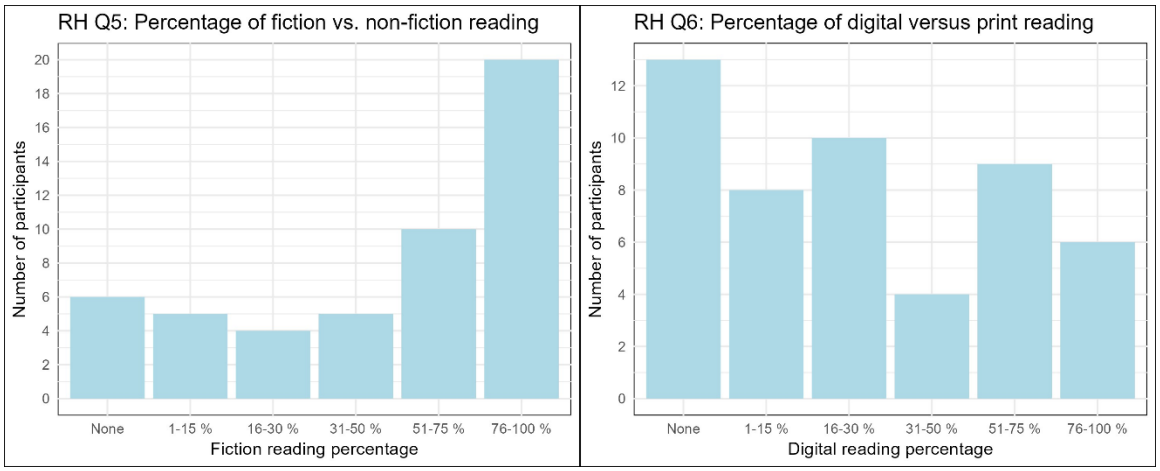


Figure 2.6: Additional reading habits survey questions which contributed to the revised reading habits questionnaire. (Left:) "Percentage of fiction vs. non-fiction reading" (Right:) "Percentage of digital versus print reading".

Motivation

Figure 2.7 displays the distribution of participants' motivation scores ($M = 3.62$, $SD = 0.55$), revealing a normal distribution (Shapiro-Wilk test, $p = .47$) with a very slight positive skew. The majority of scores cluster around the central range, with the highest frequency observed between scores 3 and 4. Extreme scores in either direction were relatively uncommon.

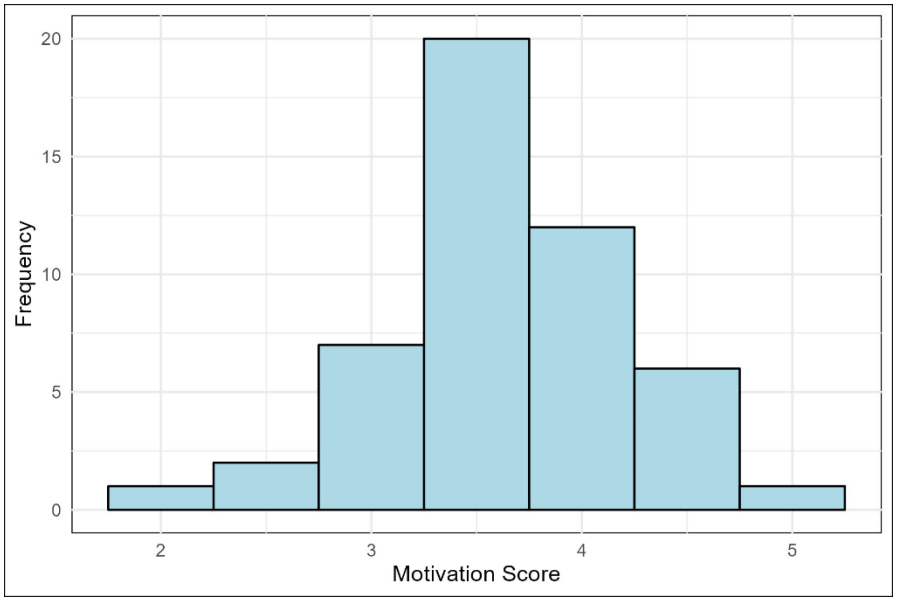


Figure 2.7: Histogram of motivation scores.

2.3.2 Statistical Analysis

Statistical analysis was conducted using R (R Core Team, 2025). Correlation tables were created using the package corrtable (Laken & Lambert, 2023). Where indicated, reliability was calculated using the alpha function from the package psych (Revelle, 2024). Linear regression models were built using the function lm from base R. Generalised linear mixed effects models (GLMER) were constructed using the package lme4 (Bates et al., 2022), *p*-values were extracted using the package lmerTest (Kuznetsova et al., 2020), and model assumptions of overdispersion, normality and outliers were checked using the package DHARMa (Hartig, 2022). To counter problems with multicollinearity, continuous predictors were first standardised before being entered into GLMERs, and we iteratively compared model performance with likelihood ratio tests using the maximal effects

structure justified by the design (Barr et al., 2013). Multicollinearity was checked using the function "vif" from the car package (Fox et al., 2022), and we followed threshold guidelines that suggest values for VIF > 5 are best avoided and VIF $>$ 10 indicates serious problems of multicollinearity (G. James et al., 2017; Menard, 2002; Vittinghoff et al., 2005) All models reported below had VIFs below 3.00. Model output tables were created using the package sjPlot (Lüdecke et al., 2023), plots were created using the package ggplot2 (Wickham et al., 2024) and Venn diagrams were created using the package VennDiagram (H. Chen, 2022).

For model comparisons, we used the Vuong likelihood ratio test for non-nested model selection (Vuong, 1989) using the package nonnest2 (Merkle et al., 2020), and the Akaike Information Criterion (AIC; Akaike, 1974, Burnham & Anderson, 2004) using the function AIC from base R. In cases where models differed in the number of predictors used, we used the Bayesian Information Criterion (BIC; Burnham & Anderson, 2004; Schwarz, 1978), which penalises model complexity more heavily than AIC to determine if the improvement in model fit justifies the added complexity. Information-theoretic paradigms such as AIC cannot determine significance and are not used for testing hypotheses (see Anderson & Burnham, 2002)—however, they can be informative for selecting between a number of different non-nested models which predict the same outcome variable. Higher delta values (ΔAIC) relative to the preferred model indicate the degree of information loss in selecting these lesser models. Models with delta values of >7 have far less support, whereas >10 indicates a model which is wholly unsupported (Burnham & Anderson, 2004). Conversely, a lower AIC value indicates better model fit. Accordingly, we use AIC primarily as a potentially more sensitive indicator of

model fit and precision than a more conservative estimate like the Vuong test, but do not rely on it to test hypotheses.

Below, we begin by showing Spearman correlations between all measures of interest (Table 2.3.9). In general, correlations were moderate to high, although AFT and ART scores were only moderately correlated ($r = .40$).

Table 2.3.9: Spearman correlation (ρ) table for all predictor and outcome variables of interest.

	AFT	*ART*	*AFAR*	*LDT (%)*	*LDT RT*	*RH*	*RHR*	*Motivation*
AFT	1							
ART	**.397****	1						
AFAR	**.547*****	**.936*****	1					
LDT (%)	**.391****	**.606*****	**.685*****	1				
LDT RT	-.106	-.087	-.128	-.047	1			
RH	**.431****	**.289***	**.360***	**.370****	-.066	1		
RHR	**.537*****	**.481*****	**.493*****	**.453*****	-.063	**.688*****	1	
Motivation	.176	**.440****	**.414****	**.507*****	-.053	.129	.179	1

LDT (%) = accuracy score for both words and non-words. *LDT RT* = response times for words only. *RH* = reading habits survey scores. *RHR* = reading habits (revised) scores. Significant correlations are in **bold**; * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

H1: Predicting Reading Habits Scores

Our first hypothesis (H1) was that AFT would be significantly more positively correlated with reading habits scores than ART. We also wanted to examine the potential effects of motivation on the outcome, and to quantify the extent to which

the two print exposure measures make different predictions for the outcome variable. To do this, we first built separate linear regression models using AFT and ART predicting reading habits composite scores, using the original items from our pre-registration. The model with AFT as sole predictor performed well with a large effect size, $F(1, 48) = 23.64$, Adj-$R^2 = .32$, $p < .001$. An analogous model using ART fared more poorly, with a small to medium effect size, $F(1, 48) = 5.80$, Adj-$R^2 = .09$, $p < .05$. Motivation was a non-significant predictor and was excluded from these models. Model fit comparisons using AIC showed a substantial preference for the AFT model, ΔAIC = -14.31. On an exploratory basis, we also built a model using the composite Author Fluency and Recognition (AFAR) metric, which combines unique and valid names from both tests into a single score. This model performed similarly to the ART model, with a medium effect size, $F(1, 48) = 9.10$, Adj-$R^2 = .14$, $p < .01$. Although the AFAR model was preferred to the ART model (ΔAIC = -2.98), the AFT model was preferred to the AFAR model (ΔAIC = -11.34). Despite the relatively improved model fit, Vuong tests showed only a marginal preference for the AFT model over ART ($z = 1.50$, $p = .07$) and AFAR ($z = 1.31$, $p = .09$) models. Since AIC values cannot be used for hypothesis testing and Vuong tests were equivocal, we conclude that with respect to the original reading habits questionnaire, we fail to reject the null hypothesis, finding only marginal evidence for H1.

As mentioned previously however, the reliability of the reading habits measure was low, and we devised an exploratory revised measure of reading habits scores by substituting some of the original questions from the reading survey with others. We then built models predicting these revised reading habits scores (RHR:

Reading Habits Revised) using our print exposure measures, and the outcomes were virtually identical to those described above. Firstly, the model with AFT as sole predictor performed well with a large effect size, $F(1, 48) = 24.93$, Adj-$R^2 = .33$, $p < .001$; full model output is provided in Table 2.3.10, and results are visualised in Figure 2.8. An analogous model using ART fared more poorly, with a small to medium effect size, $F(1, 48) = 12.74$, Adj-$R^2 = .19$, $p < .001$. Again, motivation was a non-significant predictor. Model comparisons using AIC showed a substantial preference for the AFT model over the ART model (ΔAIC = -9.15). As with the original version of the reading habits scores, we again built a model using AFAR as the sole predictor, which performed similarly to ART, with a medium effect size, $F(1, 48) = 14.46$, Adj-$R^2 = .22$, $p < .001$. Although the AFAR model was slightly preferred to the ART model (ΔAIC = -1.40), the AFT model was preferred to the AFAR model (ΔAIC = -7.75). As before, Vuong tests showed no overwhelming preference for either the AFT, ART, or AFAR models (all ps $> .05$). We conclude that although print exposure correlates with reading habits scores, there was no evidence that AFT was a superior measure to ART, disconfirming H1.

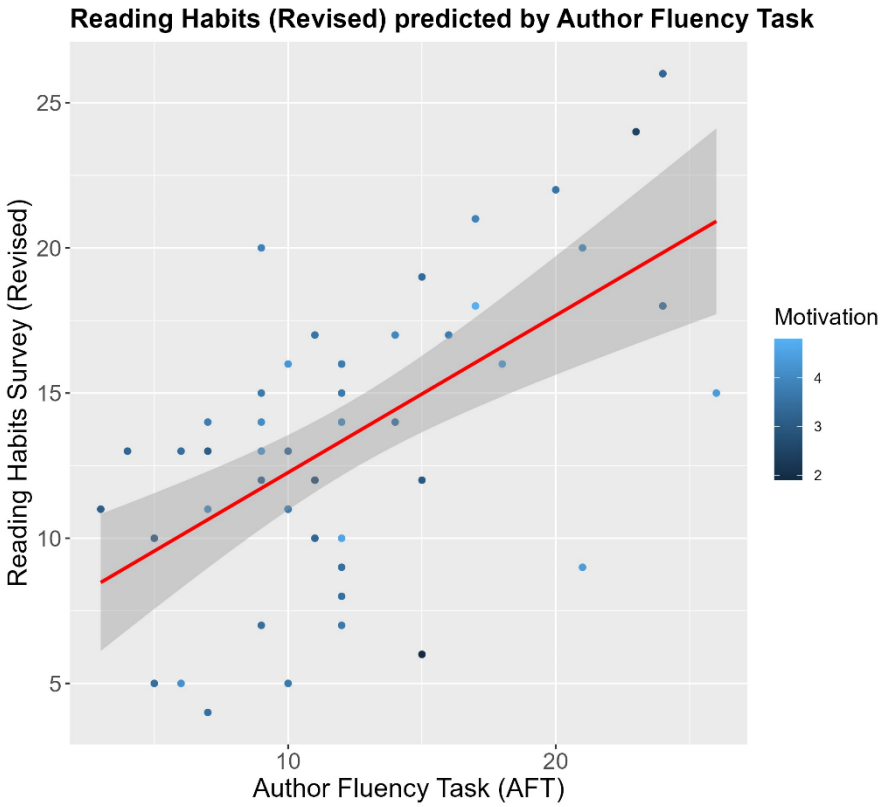


Figure 2.8: Line of best fit predicting reading habits revised (RHR) survey scores by the Author Fluency Task (AFT). Blue dots show individual observations, with lighter-coloured dots indicating higher motivation (non-significant).

Table 2.3.10: Regression output for a model predicting reading habits (revised) scores as a function of AFT.

	Reading Habits (Revised) ~ AFT		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	6.854	3.913 – 9.796	$<$ **0.001*** **
AFT	0.541	0.323 – 0.759	$<$ **0.001*** **
Observations	50		
R^2 / R^2 adjusted	0.342 / 0.328		

In summary, these results indicate that although the Author Fluency Task was a better fit for the data compared the Author Recognition Test as a predictor of reading habits scores, statistical tests were equivocal. This effectively disconfirmed H1, which posited that AFT would be more highly associated with reading habits than ART. Surprisingly, the composite AFAR measure is also not as effective (lower R^2, higher ΔAIC) a predictor as AFT alone.

Despite this, if we assume that respondents are honestly and accurately assessing their personal reading habits, our findings provide the first evidence of convergent validity for AFT as a predictor of print exposure. However, as detailed extensively elsewhere, ART was designed as a response to the assumption that such self-report measures are potentially unreliable. In the next section, we evaluate how AFT performs with a more objective measure of reading experience, using a lexical decision task with keywords of literary fiction.

H2: Lexical Decision Accuracy

Our second hypothesis (H2) was that AFT would be significantly more positively correlated with lexical decision accuracy scores compared to ART. Once again, the correlation matrix in Table 2.3.9 might suggest an obvious conclusion on its own, with ART in fact far more highly correlated with LDT scores ($r = .61$) than AFT ($r = .39$). As before however, we also wanted to account for motivation, and to compare model performance more precisely. Firstly, we normalised our predictors in all models to account for the varying scales used. Base models with predictors of either AFT or ART were both improved by adding motivation as a covariate. The model with ART and motivation showed effects of ART, $F(1, 47) = 31.95$, $p < .001$; and motivation, $F(1, 47) = 8.76$, $p < .01$; model $F(2,47) = 20.36$, Adj-$R^2 = .44$, $p < .001$; and was slightly preferred (ΔAIC = -4.14) to the analogous AFT model which showed effects of AFT, $F(1, 47) = 14.33$, $p < .001$; and motivation, $F(1, 47) = 19.42$, $p < .001$; model $F(2,47) = 16.87$, Adj-$R^2 = .39$, $p < .001$. Models with an interaction term for motivation were non-significant. However, Vuong tests comparing models with the two measures showed no preference for either model ($p > .05$), likely due to the influence of the motivation covariate, which also correlated moderately with ART (see Table 2.3.9). For comparison, excluding the motivation covariate from both models showed a marginal preference for the ART only model (Vuong test $p = .06$). This also improved the AIC preference for ART (ΔAIC = -12.89). This relative superiority of the ART model disconfirmed H2, which anticipated that AFT would outperform as a predictor of lexical decision accuracy.

As an exploratory step, we also evaluated the performance of the AFAR measure as a predictor of LDT accuracy. This model was significantly preferred

over both the ART (Vuong z = -2.00, p = .02; ΔAIC = -9.58) and AFT models (Vuong z = -1.70, p = .04; ΔAIC = -13.71), with effects of AFAR, $F(1, 47) = 50.45$, $p < .001$; and motivation, $F(1, 47) = 8.78$, $p < .01$; model Adj-R^2 = .54, $p < .001$. For comparison, Vuong and AIC tests comparing models excluding the motivation covariate showed a significant preference for the model with AFAR over both ART (Vuong z = -1.73, p = .04; ΔAIC = -9.56), and AFT (Vuong z = -2.80, p = .003; ΔAIC = -22.44).

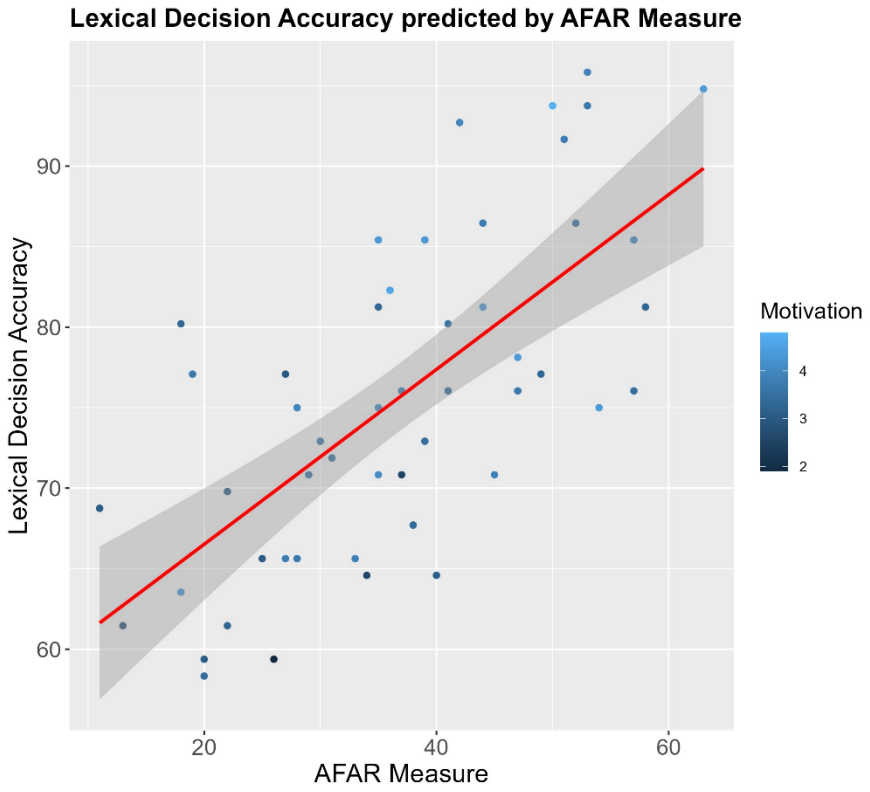


Figure 2.9: Line of best fit predicting lexical decision accuracy scores (all stimuli) as a function of Author Fluency and Recognition (AFAR) scores. Dots show

individual observations, with lighter-coloured dots indicating higher motivation (significant, $p < .01$).

Table 2.3.11: Regression output for a model predicting lexical decision accuracy scores as a function of AFAR and motivation. Predictors were normalised prior to entry in the model.

	LDT Accuracy ~ AFAR + Motivation		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	75.583	73.635 – 77.532	**< 0.001*** **
AFAR	5.687	3.541 – 7.834	**< 0.001*** **
Motivation	3.162	1.015 – 5.309	**0.005****
Observations	50		
R^2 / R^2 adjusted	0.558 / 0.539		

Generalised linear mixed effects (GLME) models were also constructed to account for random effects. These modelled the probability of each participant correctly responding to each word or non-word trial on the lexical decision task. For stimulus type, pseudowords were treatment (dummy) coded as the reference level. To account for the different scales used, we first standardised our continuous predictors. Given the superior performance of AFAR for the fixed-effects only regression model, we focus our analysis on this predictor below.

Predicting the probability of correct selections to LDT stimuli per trial, the base model with random intercepts for items was significantly improved by adding by-participant random intercepts and slopes for stimulus type (i.e. non-word/word), as well as additional fixed effects of motivation scores and an interaction between AFAR and stimulus type. Significant main effects were found both for stimulus type [word] (odds ratio [OR] = 0.06, 95% CI = [0.03, 0.12], $p <$ .001), showing participants were significantly (though very slightly) poorer at accurately classifying existing words compared to non-words; and motivation (OR = 1.29, 95% CI = [1.11, 1.51], $p <$ .01), showing more highly motivated participants performed better; but no main effect was seen for AFAR. However, the interaction between AFAR and stimulus type showed that the beneficial effect of AFAR was increased for responses to real words compared to non-words (OR = 1.95, 95% CI = [1.21, 3.14], $p = <$ .01). Specifically, for every 1 standard deviation (SD) increase in AFAR scores (equivalent to 5.49 author names; see Table 2.3.4), compared to non-word trials, the odds of correctly identifying real words were 95% higher, suggesting that reading experience was strongly associated with recognition of these keywords of fiction. We attempted to include covariates for word prevalence and fiction keyness, but these were non-significant. Likelihood ratio tests between our best model and a model with an additional interaction for motivation on AFAR was only marginally significant, and explained little additional variance. Moreover, the ΔBIC value was -4.8 in favour of the simpler model, suggesting the added complexity was not worth the slight increase in explained variance. In the interest of parsimony, then, we provide the model shown below.

The model output indicated that random effects accounted for a substantial portion of the variance, with fixed effects explaining 30.2% and the full model with random effects included explaining 57.1%. Random effect variance components showed substantial variability between items ($\tau 00 = 1.31$) and participants ($\tau 00 = 0.93$), and the ICC indicated that 39% of the variance in responses was attributable to these random effects. The full mixed effects output is shown in Table 2.3.12, and fixed effects are illustrated in Figure 2.10 as a forest plot.

Table 2.3.12: GLME output, predicting odds ratios [OR] of correct responses to LDT stimuli, using fiction keywords and matched pseudowords.

	GLME Model		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	29.421	17.930 – 48.279	**< 0.001*** **
AFAR	1.002	0.721 – 1.392	0.992
Stimulus Type [word]	0.061	0.031 – 0.122	**< 0.001*** **
Motivation	1.294	1.108 – 1.512	**0.001****
AFAR * Stimulus Type [word]	1.951	1.211 – 3.144	**0.006****
Random Effects			
σ^2	3.29		
τ_{00} Items	1.31		
τ_{00} Participants	0.93		
τ_{11} Participants.Stimulus Type [non-word]	2.38		
ρ_{01} Participants	-0.91		
ICC	0.39		
N Items	96		
N Participants	50		
Observations	4800		
Marginal R^2 / Conditional R^2	0.302 / 0.571		

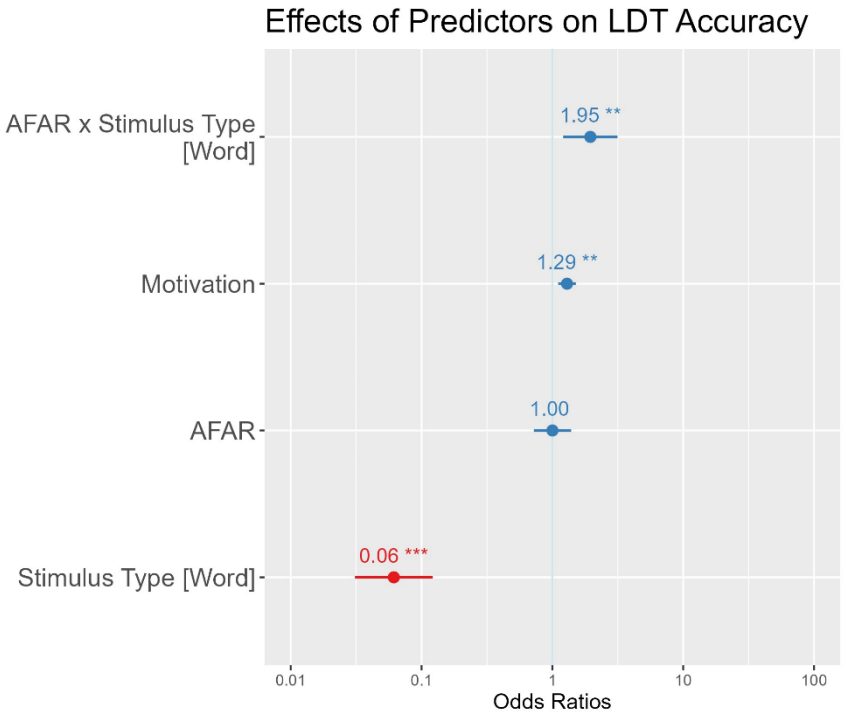


Figure 2.10: Forest plot showing odds ratios for fixed effects of AFAR, motivation, and stimulus type, as well as the interaction of AFAR on stimulus type, on the probability of correct responses on LDT. Horizontal lines indicate the confidence intervals for each effect. Red lines represent reduced odds of correct responses, blue lines represent increased odds.

We evaluated equivalent models which substituted AFAR with either AFT or ART using the Akaike Information Criterion (AIC). The best fitting model was the one which included AFAR, which carried 99% of the cumulative model evidence weight (see Table 2.3.13) and was strongly preferred to analogous models with ART (ΔAIC = -9.96) and AFT (ΔAIC = -13.74).

Table 2.3.13: AIC model comparisons; lower AIC is better. Results indicate AFAR is the best-fitting model. "AIC weight" is the cumulative model weight.

Model	K	AIC	Delta AIC	AIC weight	log-Likelihood
AFAR	9	3819.05	0.00	0.99	-1900.50
ART	9	3829.01	9.96	0.01	-1905.49
AFT	9	3832.79	13.74	0.00	-1907.37

Taken together, these analyses indicate that the combined AFAR measure explains more variance in lexical decision performance than either AFT or ART alone.

2.4 General Discussion

This pilot study compared a semantic fluency measure of print exposure using author names (AFT) to a standard ART as an associate of reading habits and vocabulary knowledge. To assess vocabulary knowledge, we used a novel lexical decision task using keywords of literary fiction and scored for accuracy on both words and non-words. As a reminder, H1 posited that although AFT and ART would both predict reading habits scores, a model with AFT as the sole measure of print exposure would outperform analogous models with ART. H2 hypothesised that both measures would also predict LDT accuracy, but AFT models would similarly outperform those with ART.

With respect to these hypotheses, H1 was disconfirmed, as although AFT was more positively correlated with self-reported reading habits than ART, this effect was not significant. H2 was similarly disconfirmed, with evidence showing

that ART was in fact marginally preferred over AFT for lexical decision accuracy. However, these results demonstrate for the first time that an open-ended fluency task for author names (AFT) performs very similarly to an ART as a predictor of print exposure, suggesting it may provide a useful alternative in some cases. Finally, exploratory analyses also showed that the novel AFAR measure, combining author names from both ART and AFT, was a better predictor of lexical decision performance than either individual predictor alone.

Perhaps more importantly than which measure is superior to the other, these findings suggest that although AFT and ART both ostensibly measure print exposure, they likely capture different aspects of this experience. Obviously, the two tasks both use author names as the fundamental unit of measurement, but the methods are distinguished through recognition versus recall memory. Differences in the kinds of memory required for each task may also partially explain why AFT is more highly correlated with the reading habits survey, which also requires participants to draw on their explicit recall of reading experience, whereas ART is more highly correlated with a recognition judgement measure like the lexical decision task. This may be even more relevant for this particular study, as the ART was carried out online and followed the same presentation as LDT, with names and foils presented serially and randomly. Consequently, the choice of outcome measure and the nature of its response mode may also be an important consideration.

Moreover, the variety of author names on AFT (and the high number of authors who were named just once) suggest there is substantial variability in the nature and amount of reading experience of participants. Given the relatively uniform nature of this cohort (mainly Oxford undergraduates, L1 UK English

speakers, mostly women under the age of 22), this is somewhat surprising. Further, it calls into question whether an ART alone, with its restricted set of author names reflecting a restricted kind of reading experience, captures the true diversity of exposure to print in an L1 population, let alone in L2. Even without capturing this diversity, however, we recognise that ART clearly works well on its own as a correlate of vocabulary skill in this sample.

Additionally, we found exploratory evidence that author name selections on AFT likely reflect demographic characteristics of participants themselves, with male respondents significantly more likely to list male authors, as well as authors writing primarily works of non-fiction. Importantly, these biases suggest the author names provided on AFT are more likely to be indicative of personal or "primary" print exposure (Martin-Chang & Gould, 2008), or actual reading experience as opposed to general cultural knowledge about literature or writers.

Although the results are promising, we recognise that the study recruited from a convenience pool of L1 English Oxford University students. Print exposure tests such as ART are typically validated in similar populations, and unsurprisingly, both ART and its derivative measures (AFT and AFAR) also correlated well with the outcome measures in this study. Yet as detailed in the introduction to this thesis, ART is not always useful in populations with lower levels of education, nor among individuals who speak English as L2. Accordingly, the present data are not generalisable beyond this study, and AFT will require extensive testing in diverse groups.

One of the fundamental reasons for using proxy measures such as ART is that self-report measures of reading habits are considered to be unreliable. Here, our initial selection of reading habits questions did indeed return poor reliability—however, the revised reading habits questionnaire was adequate. Moreover, it correlated moderately with all related measures (see Table 2.3.9), which speaks to its convergent validity. Thus, the assumption that participants cannot be trusted to respond accurately to subjective surveys may not always be well-founded. Even so, the use of such surveys relies on careful item selection and validation—in this case, we simply hand-selected questions from other reading habits surveys which we anticipated would likely be associated with print exposure, rather than using an "off-the-shelf" measure. The low reliability of our initial composite reading habits measure suggests this may not have been the right approach, yet regression models predicting the original reading habits composite scores as well as the revised scores from our print exposure measures were virtually identical. Evidently, whatever impact this low reliability may have had, it was not reflected in our models.

Another consideration is that by its nature, an LDT is essentially a measure of word recognition and single-word vocabulary knowledge. While vocabulary is naturally of interest for gauging print exposure, and is generally a reliable proxy itself for other language skills, considered alone it is insufficient to make broader generalisations about language ability, or the extent of individual reading experience. Future studies should include a more varied battery of tests of other component skills of reading, perhaps more directly targeting difficulties which L2 English speakers may encounter during reading. One possibility of this kind which

will be explored in subsequent chapters is the "chunking" together of formulaic phrases, which requires the integration of multiple words into a larger context.

Although the novel LDT used in this study indicates that keywords of literary fiction can be used as a correlate for print exposure, we also note that performance may differ on non-fiction keywords. Naturally, individuals with greater print exposure may have better vocabulary knowledge in general, yet it is unclear if this is due to additional fiction reading in particular. It may be the case, for example, that keywords of fiction are a better indicator of reading experience than keywords of non-fiction. Such a hypothetical effect would arguably reflect the relative importance of fiction reading for developing language skill, a finding reported extensively in the literature (and summarised in the introduction to this thesis, see section 1.1.1). Future studies of this kind may wish to include non-fiction keywords which are matched in prevalence and frequency to address this question, or may consider using the degree of an individual's fiction or non-fiction preference (obtained from a reading habits survey) as a covariate in these models.

Noting the relatively low mean accuracy scores for fiction keywords (61.12%, see Table 2.3.1), it is also possible that the words in the lexical decision task are simply too difficult for most participants, which may explain why reaction times were not predicted by the print exposure measures. The difficulty of this task may become even more evident when expanding to a more diverse cohort of L2 participants, who generally have less reading experience, and thus fewer potential encounters with these lower-prevalence words. The addition of some more commonly known words may therefore be necessary to capture this variation in L2 experience.

Finally, at first glance, the present results may appear to hold limited theoretical interest—ultimately, as another proxy measure for print exposure, it may be argued that AFT merely offers a variation on the same logic as ART, and it doesn't provide new insight or implications into the factors leading to improved reading skill. Moreover, although the combined AFAR measure performed best for predicting LDT accuracy, this measure is fundamentally a supplement to ART—yet a more interesting approach might be to determine how explicit and implicit measures of print exposure differ in their correlations with related skills. One way of doing this would be to select vocabulary outcome measures which differ in the degree to which they tap into explicit and implicit memory, an approach which we detail in subsequent chapters. However, we argue that the primary contribution of the present research is in demonstrating that a semantic fluency task for author names shows convergent validity with similar measures—as well as divergent validity for dissimilar ones (e.g., task motivation)—and serves as a promising proof-of-concept for adapting AFT to an L2 context, where practical and theoretical questions about how to assess print exposure may be more relevant.

2.4.1 Conclusion

Print exposure, as measured by both an Author Fluency Task (AFT) and an Author Recognition Test (ART), was significantly and positively associated with the recognition of keywords of literary fiction on a lexical decision task, as well as with scores from subjective reading habits questionnaires. Exploratory analyses also showed that names provided on AFT reflected established group differences in

reading habits trends between men and women, suggesting the measure may be more reflective of personal reading experience.

Although both of our main print exposure measures were associated with our outcome measures, we did not find any conclusive evidence that one or the other provided a significantly better fit to the data (although ART was marginally preferred for predicting LDT accuracy compared to AFT). We conclude that AFT may be considered a reasonable alternative to an ART in L1 English university populations, with its principal advantage being that it does not require researchers to regularly update a list of author names in response to changing reader interests.

Finally, we note that there may still be a benefit to administering both measures and combining scores from each. The significant improvement in model fit for the combined Author Fluency and Recognition (AFAR) measure as a predictor of lexical decision scores indicates that individual differences in reading experience are more varied than a closed format proxy measure like ART alone can account for, and AFT may provide an important supplement. Given the even greater variability in L2 exposure, we suspect that a measure which is more sensitive to this variation may have particularly useful practical applications in some L2 populations. In the following chapters, this possibility is explored through experimental methods exploring the links between print exposure and knowledge of formulaic language in L2.

2.5 Data Availability

Data and code used in analyses are available on OSF (https://osf.io/y8926/).

3. Semantic fluency for authors as a proxy measure of print exposure in L1 French speakers of English

Abstract

Reading experience provides critical input for language learning. This is typically quantified via estimates of print exposure, such as the Author Recognition Test (ART), although it may be unreliable in L2. The previous chapter introduced the Author Fluency Task (AFT) and demonstrated it was comparable to ART as an index of print exposure in L1 speakers, with both measures equally associated with reading habits surveys and recognition of keywords of literary fiction. This study extends these findings to an L2 setting by assessing knowledge of English discourse connectives and collocations among 60 bilingual French/English speakers, and a comparison sample of 60 L1 English speakers. L2 participants completed AFT, ART, and LexTALE in both languages. Analysis of L2 measures showed AFT more accurately predicted L2 vocabulary knowledge than ART, even when controlling for proficiency (LexTALE). Conversely, ART was more effective for L1 speakers, showing a striking dissociation between the measures across language groups. Additionally, within the French group, data showed a limited role for L1 proficiency and print exposure on advanced L2 vocabulary acquisition. These findings recommend AFT as a valuable tool for quantifying L2 print exposure's role in language learning.

3.1 Background

3.1.1 Formulaic language in L2

Some vocabulary is especially difficult for L2 speakers to acquire and use naturally, and this difficulty may be partially related to comparatively lower exposure to print. For example, discourse connectives like "nevertheless" and "consequently" link ideas from separate sentence clauses, and are often associated with written language, in particular academic writing (Biber, 2006). These connectives encode a set of instructions for interpreting the relationship between ideas, rather than having a strict lexical definition (Van Silfhout et al., 2015; Zufferey et al., 2015), and this may render them difficult to acquire through explicit teaching. One explicit approach to learning L2 connectives is simply to provide an approximate L1 equivalent—yet an analogous term in L1 may not necessarily encode the same relations in L2 in all cases (Zufferey & Gygax, 2017). This can be problematic for L2 speakers who often filter their L2 through the lens of their L1, particularly during the early stages of acquisition, relying on an unreliable equivalence between L1 and L2 vocabulary items (Ringbom, 2016). Consequently, connectives pose a serious challenge, with even very advanced L2 speakers often struggling to understand how and when to use them correctly (Wetzel et al., 2020).

Similarly, word collocations (e.g., higher frequency of *"weak tea"* over *"feeble tea"*) are another obstacle for L2 learners. Although these can be learned through explicit instruction, they are more challenging than learning single words (Peters, 2014, 2016), and their virtually endless number means they are arguably an

inefficient use of targeted language instruction, which mostly focuses on teaching individual words (Schmitt, 2010). However, collocations can be acquired incidentally through statistical learning from input, both in L1 and L2 (Pellicer-Sánchez, 2017; Sonbul & Schmitt, 2013; Webb et al., 2013). In effect, the more language input one encounters, the more these associations are formed *("items that are used together fuse together";* Bybee, 2007, p. 4). Accordingly, L2 speakers process L2 collocations more slowly than L1 speakers (Siyanova & Schmitt, 2008) and use fewer collocations themselves, which tend to be congruent with L1 (Granger, 1998).[15]

Whereas the importance of selecting the correct connective may be clear, the significance of collocation knowledge may be less evident. After all, what difference is there between *"raise prices"* and *"lift prices"*? If "raise" and "lift" are essentially synonymous, surely either one will serve the same purpose. But all word pairings are not created equal, and set phrases are subject to certain preferential selection constraints. Indeed, although speakers of a language may correctly infer the meaning of an idiosyncratic expression, formulaic language is processed more quickly (N. C. Ellis et al., 2008; Hallin & Van Lancker Sidtis, 2017). This formulaic preference thus ostensibly functions to ease processing burdens between communicators (Wray, 2002).

[15] Academic writing also commonly features formulaic "lexical bundles", or set phrases which serve specific functions like topic priming ("it has often been asserted that..."), support for an argument ("it has been shown that..."), or self-reference ("as we have seen above..."), in which the main verb of each phrase may be substituted for a similar or related verb, but which are otherwise essentially fixed (Nattinger & DeCarrico, 2010; Oakey, 2002). Unsurprisingly then, these are also less common in academic writing by advanced L2 speakers (N. C. Ellis, 2012; Granger, 1998).

Both connectives and collocations are examples of formulaic language, and what constitutes as "formulaic" is largely (though not solely) a matter of frequency of exposure (Siyanova-Chanturia, Conklin, & Van Heuven, 2011). However, an expression's frequency is dependent on modality and input type, with corpus studies revealing that certain collocations are more common in written or spoken language (Gablasova et al., 2017; D. Shin, 2007), connectives use is more varied in writing compared to speech (Tskhovrebova et al., 2022), and connective frequencies vary by modality and register (M. Andersson & Sundberg, 2021). Given that learners have comparatively lower L2 exposure, and tend to interpret formulaic language in L2 serially (i.e. word-by-word) rather than processing into meaningful "chunks" as in L1 (Conklin & Schmitt, 2012), connectives and collocations present a significant hurdle. Accordingly, L2 writing and speech is often characterised by an overreliance on certain connectives (Wetzel et al., 2020), and features less formulaic language in general (Granger, 1998; Pérez-Llantada, 2014). Formulaic expressions, however, are often less about the meaning of individual words than understanding how words relate to each other. As J.R. Firth put it, echoing Wittgenstein, *"you shall know a word by the company it keeps"* (1957, p. 11).

3.1.2 Contributions of L1- and L2-specific skills for L2 learning

Although the importance of L1 input is well-accepted, the degree of influence of L1- vs L2-specific skills in second language acquisition (SLA) remains a matter of debate. Language transfer theories (Baker et al., 2011; Cummins, 1979; Sparks,

1995) posit that greater L1 proficiency affords the potential for greater proficiency in L2, and while there is considerable evidence for this (e.g., Berthele & Vanhove, 2020), some have argued it is limited to more general language skills such as phonology and pragmatics rather than syntax and vocabulary (Verhoeven, 1994). Yet some evidence contradicts this view of language transfer being fundamental for L2 development. Notably, a recently updated meta-analysis shows L2 reading comprehension is determined primarily by L2 skills, even above language transfer or general cognitive skills (Jeon & Yamashita, 2014, 2022). Additionally, in a massive study of English learners from different backgrounds on various component skills of reading, Siegelman et al. (2023) demonstrate that L1/L2 group differences are often minimal, arguing the differences observed between language groups are partly due to the nature of hypothesis testing. The authors propose that the range of ability should be considered one which encompasses both L1 and L2 speakers, and they emphasise the significant degree of overlap in skills. This is more consistent with a meta-analysis which argued in support of interdependence and contrastive theories (Melby-Lervåg & Lervåg, 2011).

For our present discussion, the role of L1 print exposure is particularly relevant, and there is evidence for its influence on L2 reading skills, including decoding and comprehension (Sparks et al., 2012). One study showed that while L1 German print exposure (as measured by a German ART) predicted L2 French connectives knowledge, a French ART did not (Wetzel et al., 2020). Although this may be attributable to language interdependence, we contend that the findings are expected for this population of adolescent beginner L2 speakers, who generally have little L2 exposure—as the authors point out, their participants knew very few of

the second-language authors on ART. Since the effect of print exposure is cumulative over a lifetime, a more interesting case might be to compare L1 and L2 print exposure measures in a proficient L2 population. This is what we endeavoured to do in the present study.

3.2 Present Study

This study received ethics approval [reference R77364/RE002] and was pre-registered (https://osf.io/nsduz/). We tested whether L1 French / L2 English reading experience (assessed by AFT and ART in both languages) is associated with individual differences in knowledge of English connectives and collocations, even when accounting for a standard proficiency measure in both languages. Our research questions were intended to assess the utility of AFT as a novel measure of print exposure:

1) Does an L2 AFT outperform ART as a predictor of advanced L2 vocabulary knowledge? Does either measure explain additional variance not accounted for by proficiency?
2) Do AFT/ART perform differently by vocabulary measure (collocations vs. connectives)?
3) Does L1 or L2 print exposure better predict performance on L2 vocabulary tasks?

For our L2 English cohort, we hypothesised that:

1) L1/L2 LexTALE scores would both positively predict connectives and collocations scores.
2) L2 (but not L1) AFT scores would positively predict connectives scores when controlling for LexTALE.
3) L1/L2 ART (but not AFT) scores would positively predict collocations scores when controlling for LexTALE.

For comparison, we hypothesised the same pattern for our L1 English cohort, i.e. that the English ART would predict collocations scores, and AFT would predict connectives. Essentially, we predicted that L2 print exposure, measured by AFT, would reliably predict connectives scores when controlling for LexTALE, but only ART would predict additional variance for collocations scores. The rationale for this prediction was that as an implicit judgement task, ART may recruit similar skills as those required for the collocations task. In contrast, we posited that the L2 English AFT, reflecting explicit memory of L2 reading experience, would be associated with English connectives scores even when controlling for LexTALE. The L1 French AFT, however, as an index of explicit memory of L1 reading, was not anticipated to predict L2 vocabulary. In this way, we aimed to determine how L1 and L2 print exposure variously contribute to L2 language skills.

3.3 Methods

3.3.1 Participants

Prior to data collection, power analysis was carried out using G*Power (Faul et al., 2007). For 0.8 power to detect a small effect size of .15 at .05 alpha error probability, we obtained a recommended sample size of $n = 55$.

Sixty L1 French/L2 English participants (M_{age} = 31.13, 32 female) were recruited through Prolific (2024) to complete a single session on the online experimental research platform Gorilla (Anwyl-Irvine et al., 2020). Participants who provided informed consent and completed the study were reimbursed £6.67 each. Selection criteria required participants be between 18-75 years old native French speakers currently living in France, who spoke and read English fluently at an intermediate-to-advanced level, with normal or corrected-to-normal vision.

We also recruited 60 L1 English speakers (M_{age} = 39.42, 37 female) through Prolific. Selection criteria mirrored that of the L2 group, but with native English speakers living in the UK. Below, we primarily restrict our analyses to the L2 cohort, permitting us to compare the relative contributions of L1 and L2 measures. However, we also include models from L1 English speakers to illustrate the differential predictions made by our print exposure measures.

3.3.2 Procedure

Participants began by completing the 1) demographics questionnaire, followed by 2) the English and French AFT, 3) the English and French LexTALEs and Author

Recognition Tests, 4) the Connectives and Collocations tasks, and 5) the motivation survey. Task order was counterbalanced for levels 2, 3, and 4 due to task similarities. The L1 participant procedure was identical, excluding French tasks. Figure 3.1 illustrates the task sequence.

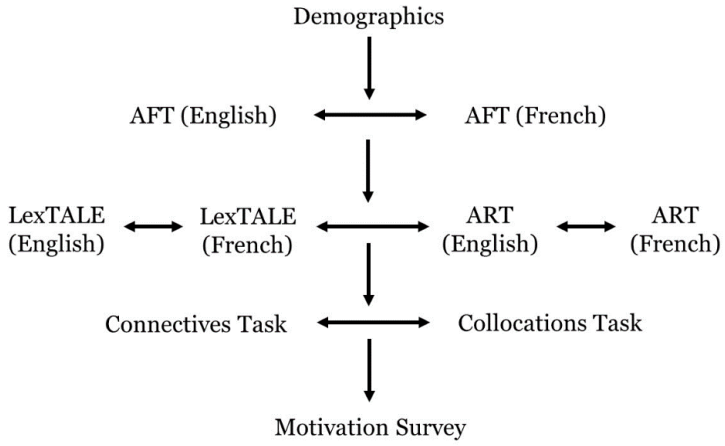


Figure 3.1: Task sequence for L2 speakers. Bidirectional arrows indicate counterbalanced orders due to task similarities. The L1 participant procedure was identical, excluding French tasks.

3.3.3 Measures

AUTHOR FLUENCY TASK

The Author Fluency Task was the same as the version described in Chapter 2. The French AFT was identical in procedure and scoring, but with French-language instructions. Accordingly, participants were asked to provide names of authors who have been published in French. The full instructions can be found in Appendix A.1.

AUTHOR RECOGNITION TEST

The English Author Recognition Test was the same as the version described in Chapter 2. The French ART followed the same procedure and scoring logic, but participants were provided instructions in French. This version was taken from Zufferey and Gygax (2020), and featured 40 author names and 40 foil names. Of the 40 authors included on the French ART, 20 were authors who wrote or were published primarily in French, whereas the other half were foreign writers. The full list of author names and foils is provided in Appendix A.7.

LEXTALE

The English LexTALE (Lemhöfer & Broersma, 2012) is a lexical decision task containing 40 words, 20 non-words, and 3 filler words. Participants were randomly presented with each item and were asked to indicate which they recognised as English words. Performance was assessed as the percentage of correct selections for words and non-words out of the total.

The French LexTALE (Brysbaert, 2013) followed the same procedure and scoring logic, but contained 56 French words and 28 non-words. Instructions were provided in French, and performance was assessed as the percentage of correct selections out of the total.

DISCOURSE CONNECTIVES TASK

This task was adapted and translated to English from the original version in Wetzel et al. (2020), which was presented in French to L1 German French learners. This is a sentence cloze task which asks participants to complete a sentence by selecting the appropriate connective from six options. For example:

Nadir likes his job, _____ his sister

would like to change her career.

1) therefore 2) since **3) whereas**

4) furthermore 5) nevertheless 6) as long as

Each connective falls into one of six "coherence relations" denoting the logical relationships specified by each connective, e.g., "*whereas*" encodes a "contrast" relation. For each sentence, competitors were selected from each of the other relations. High and low frequency connectives were selected using the corpus English Web 2020 ("enTenTen20") in corpus software *SketchEngine* (Kilgarriff et al., 2014). Connectives are shown by coherence relation and frequency in Table 3.3.1. The full list of stimuli can be found in Appendix A.12.

Table 3.3.1: Frequency of connectives in the corpus English Web 2020 ("enTenTen20"), using SketchEngine (Kilgarriff et al., 2014).

Relation	Connective	Frequency	Number of occurrences	Per million words
Addition	indeed	high	3,217,209	74.60
	furthermore	low	1,143,279	26.51
Cause	since	high	20,383,701	472.66
	given that	low	560,060	12.99
Concession	despite	high	4,852,182	112.51
	nevertheless	low	983,592	22.81
Condition	as long as	high	1,497,645	34.73
	provided	low	201,774	4.68
Consequence	therefore	high	5,495,349	127.43
	hence	low	1,441,518	33.43
Contrast	whereas	high	1,337,974	31.03
	conversely	low	190,419	4.42

COLLOCATIONS TASK

The *Words That Go Together* task was used to assess knowledge of English collocations (Dąbrowska, 2014). Participants read a list of five word-pair phrases and were instructed to select the one which is most familiar or natural. Accuracy scores were calculated as percentages of correct selections. For the full list of items, see Appendix A.13.

SEMANTIC FLUENCY

After reviewing the initial findings, we contacted the participants and invited them to complete an additional task of general semantic fluency in English. This permitted an additional analysis (pre-registered in an update). The task followed the same format as AFT, but with three different categories of items: "animals", "grocery items", and "famous people" (i.e. public figures or celebrities). Of the original 60 L2 participants we invited to take part in this follow-up study two months later, 48 returned. Participants were given one minute for each category, for a total of three minutes, equivalent to AFT. Unlike AFT, they were unable to complete the task early, which may have increased the number of items provided. Items were scored by the first author and calculated as the sum of unique and valid items per category.

ADDITIONAL VARIABLES

Details on additional variables, including motivation and demographics, can be found in Appendix A: Measures.

3.4 Results

Summary statistics and sample sizes per task are presented in Table 3.4.1. As to be expected, L1 participant scores exceeded L2, most notably for AFT, LexTALE, and collocations. Extreme outliers were identified as those falling below Q1 -

1.5*IQR or above Q3 +1.5*IQR within their respective cohort on a particular task. While not preregistered, this step was taken due to very low scores for some tasks. Participants were removed from tasks for which they had outlier scores, leading to slightly lower sample sizes in some measures. Correlations for all measures in the L2 group are shown in Table 3.4.2. Author name selections for ART and AFT can be found in Appendix B.1 and Appendix B.2, respectively.

Table 3.4.1: Summary statistics for each task by language group, after trimming. Mann-Whitney U tests compare performance between groups on each measure, and *p*-values were Bonferroni corrected for multiple comparisons.

	L1	***n***	**Min**	**Med**	**Mean**	**Max**	**SD**	**IQR**	**U**	***p*-value**
AFT	EN	59	3	12	12.56	27	5.49	8.00	2474.5	**<.001 *****
	FR	60	1	8.50	8.92	22	5.32	9.00		
AFT-FR	FR	60	1	11	11.77	28	6.18	8.25		
ART	EN	60	5	25.50	27.23	53	11.18	14.75	2278	.08
	FR	60	0	20	21.35	46	10.4	15.25		
ART-FR	FR	60	-3	7	9.50	29	7.90	10.50		
Coll. (%)	EN	59	59.00	42.50	77.50	76.06	95	11.47	3136	**<.001 *****
	FR	60	60.00	20.00	47.50	48.38	82.50	17.23		
Conn. (%)	EN	59	61.67	86.67	85.28	98.33	8.68	13.33	2339	**<.05 ***
	FR	60	20.00	79.17	71.97	100	21.77	35.42		
LT (%)	EN	57	81.67	95.00	95.03	100	4.44	5	2834.5	**<.001 *****
	FR	60	60	84.17	83.11	100	11.25	20		
LT-FR (%)	FR	60	65.48	88.10	87.26	98.81	7.71	11.90		
SF	FR	47	14	34	35.57	60	11.42	16.00		

Coll. = collocations; *Conn.* = connectives; *LT* = LexTALE; *-FR* = French versions of various tasks; *SF* = Semantic Fluency Task.

Table 3.4.2: Spearman correlation matrix for all measures, L2 English cohort.

	AFT	AFT-FR	ART	ART-FR	Coll.	Conn.	LT	LT-FR	SF
AFT	1								
AFT-FR	**.686 *****	1							
ART	**.676 *****	**.536 *****	1						
ART-FR	**.439 *****	**.675 *****	**.464 *****	1					
Coll.	**.603 *****	**.486 *****	**.324 ***	**.300 ***	1				
Conn.	**.612 *****	**.400 ****	**.374 ****	.248	**.782 *****	1			
LT	**.605 *****	**.567 *****	**.357 ****	**.363 ****	**.820 *****	**.734 *****	1		
LT-FR	.152	.24	.132	**.388 ****	.097	-.07	.202	1	
SF	**.616 *****	**.467 *****	**.486 *****	-.002	**.510 *****	**.622 *****	**.530 *****	.007	1

Coll. = collocations; *Conn.* = connectives; LT = LexTALE; *FR* = French versions of various tasks; *SF* = Semantic Fluency Task.

Table 3.4.3: Spearman correlation matrix for all measures, L1 English cohort.

	AFT	ART	Coll.	Conn.	LT
AFT	1				
ART	**.646 *****	1			
Coll.	**.308 ***	**.372 ****	1		
Conn.	**.646 *****	**.373 ****	**.540 *****	1	
LT	.254	**.464 *****	**.405 ****	**.412 ****	1

Coll. = collocations; *Conn.* = connectives; LT = LexTALE.

Analysis was performed in R (version 2023.12.1, R Core Team, 2024). Generalised linear mixed effects models (GLMER) were constructed using the package lme4 (Bates et al., 2022), *p*-values were extracted using the package lmerTest (Kuznetsova et al., 2020), and model assumptions of overdispersion, normality and outliers were checked using the package DHARMa (Hartig, 2022). To counter problems with multicollinearity, continuous predictors were first standardised before being entered into GLMERs, and we iteratively compared model performance with likelihood ratio tests using the maximal effects structure justified by the design (Barr et al., 2013).

3.4.1 Connectives

We begin by describing performance by connective type and language group before considering models which demonstrate the relative strengths of each predictor for both language groups. Scores for each connective, by coherence relation, frequency and language group are presented in Table 3.4.4. Higher frequency connectives, unsurprisingly, were responded to more accurately than lower-frequency alternatives. A notable exception was "indeed", where performance was poorer compared to even the lowest frequency connective. This may be because we are specifically interested in its use as a subordinating conjunction, which would not have been uniquely captured with our search terms—although "indeed" is very common in the corpus, its use as this coherence relation is substantially lower

relative to alternative uses.[16] Curiously, L2 speakers outperformed L1 participants on "indeed", the sole exception of its kind. Performance by coherence relation and L1 group is illustrated in Figure 3.2.

Table 3.4.4: Accuracy scores as percentages per connective, by frequency (high/low) and by language group.

Relation	**Connective**	**Frequency**	**L1**		**L2**	
			Mean	*SD*	*Mean*	*SD*
Addition	indeed	high	0.52	0.50	0.58	0.49
	furthermore	low	0.83	0.37	0.76	0.43
Cause	since	high	0.91	0.29	0.75	0.43
	given that	low	0.94	0.25	0.71	0.45
Concession	despite this	high	0.87	0.34	0.80	0.40
	nevertheless	low	0.90	0.30	0.80	0.40
Condition	as long as	high	0.93	0.25	0.83	0.38
	provided	low	0.90	0.30	0.61	0.49
Consequence	therefore	high	0.90	0.30	0.79	0.41
	hence	low	0.79	0.41	0.62	0.49
Contrast	whereas	high	0.91	0.29	0.77	0.42
	conversely	low	0.85	0.36	0.62	0.49
Global	***Overall***		**0.85**	**0.35**	**0.72**	**0.45**
	High Frequency		0.84	0.37	0.75	0.43
	Low Frequency		0.87	0.34	0.69	0.46

[16] Consequently, it may seem reasonable to re-code "indeed" as a low frequency connective rather than a high one. To decide, we hand-coded a random sample of 500 instances of "indeed" on *SketchEngine,* and determined it appears as a connective in 39.8% of instances. While this is low, the proportionally adjusted value is 29.69 ppm, which is still slightly higher than "furthermore". We have opted to leave the original coding intact.

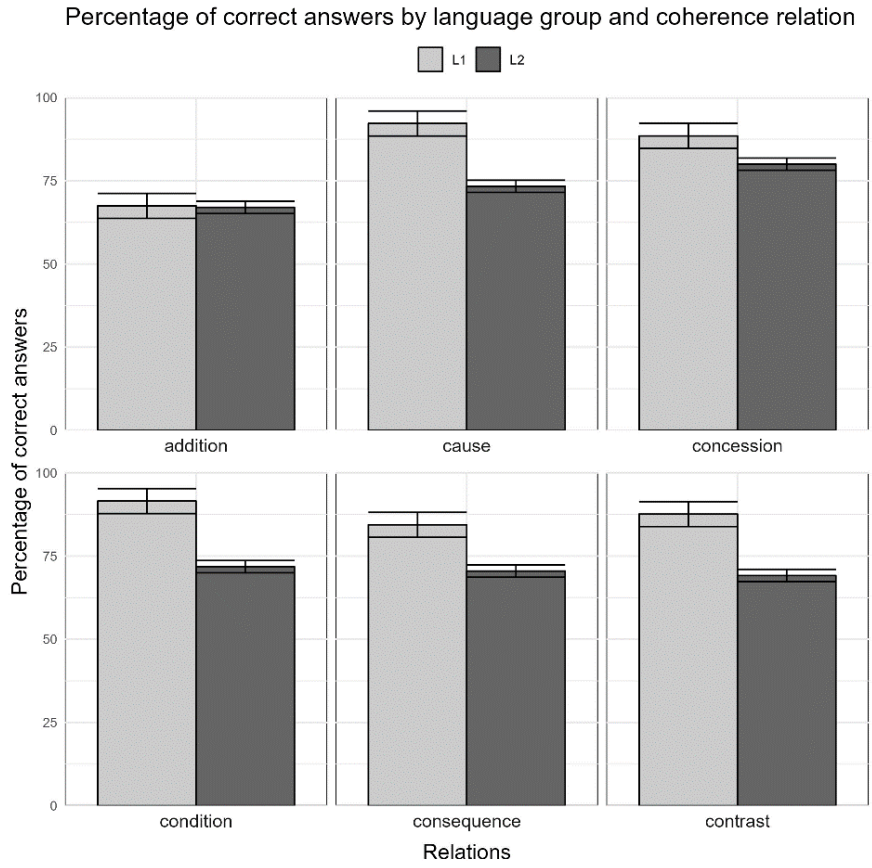


Figure 3.2: Percentage of correct answers per each coherence relation by language group (L1/L2). Error bars represent the standard error.

For L2 English speakers, comparisons favoured a regression model with only the English AFT as predictor of connectives scores over one with ART alone, as indicated by a significant Vuong test ($z = 2.54$, $p < .01$) and a lower AIC (ΔAIC = -17.76). Our most comprehensive model, $F(2, 57) = 39.86$, $p < .001$, Adj-R^2 = .57, showed effects for both LexTALE, $F(1, 57) = 73.61$, $p < .001$, and AFT, $F(1, 57) = 6.11$, $p < .05$. The English ART did not significantly predict connectives when considering either of the other variables. The contributions of each L2

predictor are illustrated in the standardised partial residuals presented in Figure 3.3 (top).

Table 3.4.5: Regression output predicting scores by L1 French participants on L2 connectives task with LexTALE and AFT.

Predictors	Estimates	CI	p
(Intercept)	-29.897	-60.617 – 0.823	0.056
LexTALE	1.109	0.693 – 1.525	**<0.001*** **
AFT	1.086	0.206 – 1.967	**0.016***
Observations	60		
R^2 / R^2 adjusted	0.583 / 0.568		

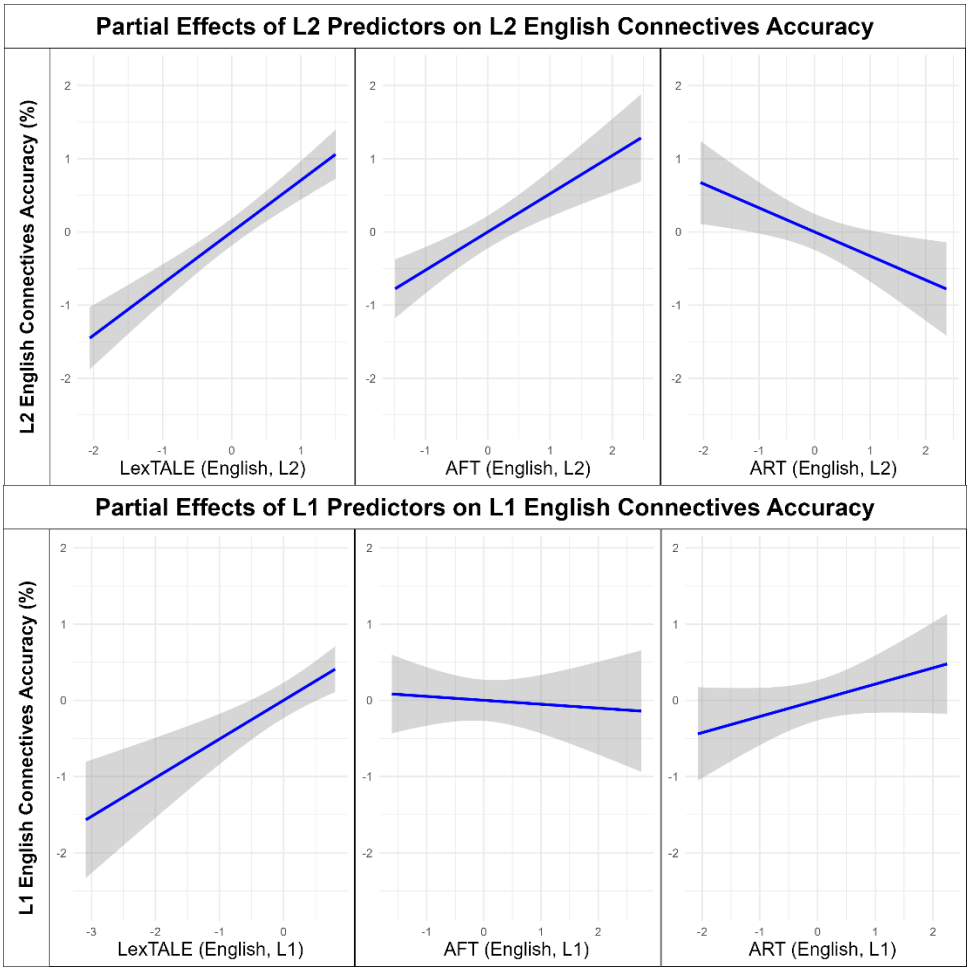


Figure 3.3: Standardised partial residual plots of all predictors on English connectives in L2 (top) and L1 (bottom). Shaded areas represent the 95% confidence interval.

Performance was also evaluated using L1 French measures. Comparing models with predictors of the French AFT and ART alone preferred the AFT model (ΔAIC = -5.20). The best-fitting model, $F(1, 58) = 9.40$, $p < .01$, Adj-R^2: .13, identified AFT-FR as a significant predictor. However, this model did not satisfy the assumption of normally distributed residuals (Shapiro-Wilk test $p < .01$) and attempts to address this issue through data transformation and robust regression

methods were unsuccessful. This model's findings are thus interpreted with caution, but evidently, the explanatory power of L1 print exposure is modest.

For L1 English speakers, separate regression models showed a modest effect of AFT on connectives, $F(1,57) = 5.97$, Adj-$R^2 = .08$, $p < .05$; whereas ART performed slightly better, $F(1,57) = 9.19$s, Adj-$R^2 = .12$, $p < .01$, and this ART only model was moderately preferred (delta AIC = -2.94). Figure 3.3 (bottom) shows standardised partial residual plots from a model with all predictors.

To illustrate the differences in the two print exposure predictors across language groups, we constructed generalised linear mixed-effects models (GLMER) as an exploratory measure. Our final model included fixed effects of AFT, ART, connective frequency and coherence relation, and their interactions with language group, as well as random intercepts for participants and items (Marginal-R^2: .16, Conditional-R^2: .34; Table 3.4.6). Contrasts were treatment (dummy) coded, with the baseline set to "low" for connective frequency, "addition" for coherence relation, and "L1" for group. Main effects for all coherence relations were significant (ORs = 2.49-5.39), though confidence intervals varied widely when comparing across groups. There was also a significant negative interaction for the L2 group for all coherence relations except for "concession". The main effect of frequency was non-significant, but interacted with language group such that L2 speakers had significantly increased odds in the high frequency condition compared to L1 speakers ($OR = 1.61$, $p < .01$). Main effects for ART and AFT were also non-significant, but there was a significant interaction between AFT and language group, such that AFT predicted increased odds ratios in L2 ($OR = 2.13$, $p < .01$). Thus, for each 1 SD increase in AFT (5.32 author names in L2), the odds of correct

selections increased by 113% for L2 compared to L1 speakers. Fixed effects are visualised in Figure 3.4.

Table 3.4.6: Fixed effects and their interactions with language group, and random effects of participant/item on odds of correct connectives selections.

Predictors	Odds Ratios	CI	p
(Intercept)	2.957	1.437 – 6.084	**.003****
ART	1.251	0.912 – 1.717	.165
AFT	1.090	0.794 – 1.496	.593
Frequency [High]	0.936	0.550 – 1.593	.809
Relation [Cause]	5.389	2.144 – 13.546	**< .001****
Relation [Concession]	3.358	1.357 – 8.309	**.009****
Relation [Condition]	4.979	1.985 – 12.487	**< .001****
Relation [Consequence]	2.494	1.014 – 6.132	**.047***
Relation [Contrast]	3.143	1.272 – 7.769	**.013***
Group [L2]	0.731	0.466 – 1.146	.172
ART x Group [L2]	0.789	0.503 – 1.236	.301
AFT x Group [L2]	2.129	1.354 – 3.348	**.001****
Frequency [High] x Group [L2]	1.611	1.225 – 2.118	**< .001****
Relation [Cause] x Group [L2]	0.282	0.176 – 0.452	**< .001****

Predictors	Odds Ratios	CI	p
Relation [Concession] x Group [L2]	0.702	0.448 – 1.101	.123
Relation [Condition] x Group [L2]	0.290	0.182 – 0.463	**< .001*** **
Relation [Consequence] x Group [L2]	0.507	0.331 – 0.777	**.002** **
Relation [Contrast] x Group [L2]	0.353	0.229 – 0.546	**< .001*** **
Random Effects			
σ^2	3.29		
$\tau_{00\ Participant}$	0.70		
$\tau_{00\ Item}$	0.18		
ICC	0.21		
$N_{\ Participant}$	115		
$N_{\ Item}$	12		
Observations	6900		
Marginal R^2 / Conditional R^2	0.158 / 0.337		

Figure 3.4: Fixed effects with interactions for language group on all variables, predicting odds of correct connectives selections. Lines represent the 95% confidence interval. Red lines represent reduced odds of correct responses, blue lines represent increased odds.

3.4.2 Collocations

We begin by describing performance generally in L2 before considering models for both language groups. In L2, trial accuracy was as low as 6.67% for *"refuse an application"* (due to competition from *"deny an application"*) to as high as 80% for *"fair share"*. Detailed statistics on the full list of items by language group are in Appendix B.5.

For L2 speakers, comparing separate, non-nested linear regression models favoured a model with AFT over one with ART (ΔAIC = -10.92), and the best fitting model, $F(2, 57) = 63.39$, $p < .001$, Adj-R^2: .68, included both LexTALE, $F(1, 57) = 123.46$, $p < .001$, and AFT, although this print exposure measure was only marginal $F(1, 57) = 3.32$, $p = .07$ (Table 3.4.7). ART was not significant when accounting for either additional variable. To illustrate the differential contributions of each predictor, standardised partial residual plots from a model with all predictors are shown in Figure 3.5 (top).

Table 3.4.7: Regression model predicting L2 English collocations scores for L1 French participants.

Predictors	Estimates	CI	p
(Intercept)	48.375	45.852 – 50.898	**<0.001*** **
LexTALE	12.360	9.164 – 15.556	**<0.001*** **
AFT	2.908	-0.288 – 6.105	0.074
Observations	60		
R^2 / R^2 adjusted	0.690 / 0.679		

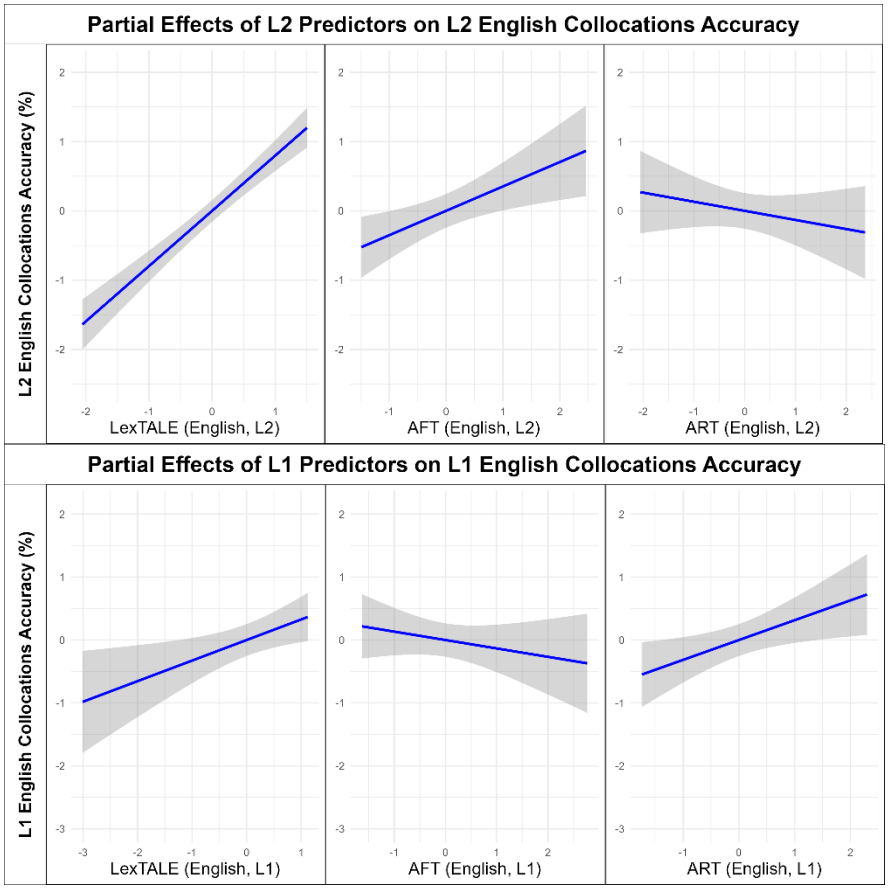


Figure 3.5: Standardised partial residual plots of all predictors on English collocations in L2 (top) and L1 (bottom).

Using L1 French predictors, separate regression models showed the French AFT, $F(1, 58) = 6.79$, $p < .05$, Adj-R^2: .09, and ART $F(1, 58) = 5.75$, $p < .05$, Adj-R^2: .07, each modestly predicted collocations scores, with negligible differences in model fit (ΔAIC = -0.97), indicating limited explanatory power for L1 print exposure. LexTALE (FR) was not associated with L2 collocations scores, demonstrating a null role for L1 proficiency.

For L1 English speakers, individual regression models predicting collocations scores showed a null effect of AFT, but a significant albeit small effect of ART, $F(1, 53) = 2.65$, $p < .05$, Adj-R^2: .10. Our best model, $F(2, 52) = 6.06$, $p < .001$, Adj-R^2: .16, included LexTALE, $F(1, 52) = 8.40$, $p < .01$, whereas ART was marginal, $F(1, 52) = 3.73$, $p = .06$. As with the connectives task, the L1 English ART was a better predictor compared to AFT—an opposite finding to L2 speakers. For comparison with L2, we provide residual plots from a model including all predictors in Figure 3.5 (bottom).

We also constructed an exploratory GLMER predicting odds of correct collocation selections. Our final model included fixed effects of AFT, ART, and collocation frequency (as a continuous measure, using values from Dąbrowska, 2014), and their interactions with language group, with random intercepts for participants and items (Marginal-R^2: .15, Conditional-R^2: .35; Table 3.4.8). Significant main effects were found for ART (OR = 1.58, $p < .001$) and language group (OR = 0.24, $p < .001$), but AFT and frequency were non-significant.

However, there were significant interactions with language group, with AFT predicting increased odds ratios in L2 compared to L1 (OR = 1.70, $p < .01$), translating into 70% higher odds per 1 SD in AFT score; and for frequency and language group, predicting increased odds ratios in L2 compared to L1 for higher-frequency collocations (OR = 1.18, $p < .05$). ART also marginally predicted lower odds in L2 compared to L1 (OR = 0.70, $p = .05$). Fixed effects are visualised in Figure 3.6.

Table 3.4.8: Fixed effects and their interactions with language group, and random effects of participant/item on odds of correct collocations selections.

Predictors	Odds Ratios	CI	p
(Intercept)	3.905	2.853 – 5.347	< **.001***
ART	1.576	1.220 – 2.035	< **.001***
AFT	0.903	0.701 – 1.162	.428
Collocation Frequency	1.052	0.808 – 1.371	.705
Group [L2]	0.243	0.185 – 0.319	< **.001***
ART x Group [L2]	0.696	0.485 – 0.999	**.049***
AFT x Group [L2]	1.701	1.189 – 2.435	**.004****
Collocation Frequency x Group [L2]	1.179	1.027 – 1.353	**.019***
Random Effects			
σ^2	3.29		
$\tau_{00\ Participant}$	0.40		
$\tau_{00\ Item}$	0.61		
ICC	0.24		
$N_{\ Participant}$	115		
$N_{\ Item}$	40		
Observations	4600		
Marginal R^2 / Conditional R^2	0.145 / 0.346		

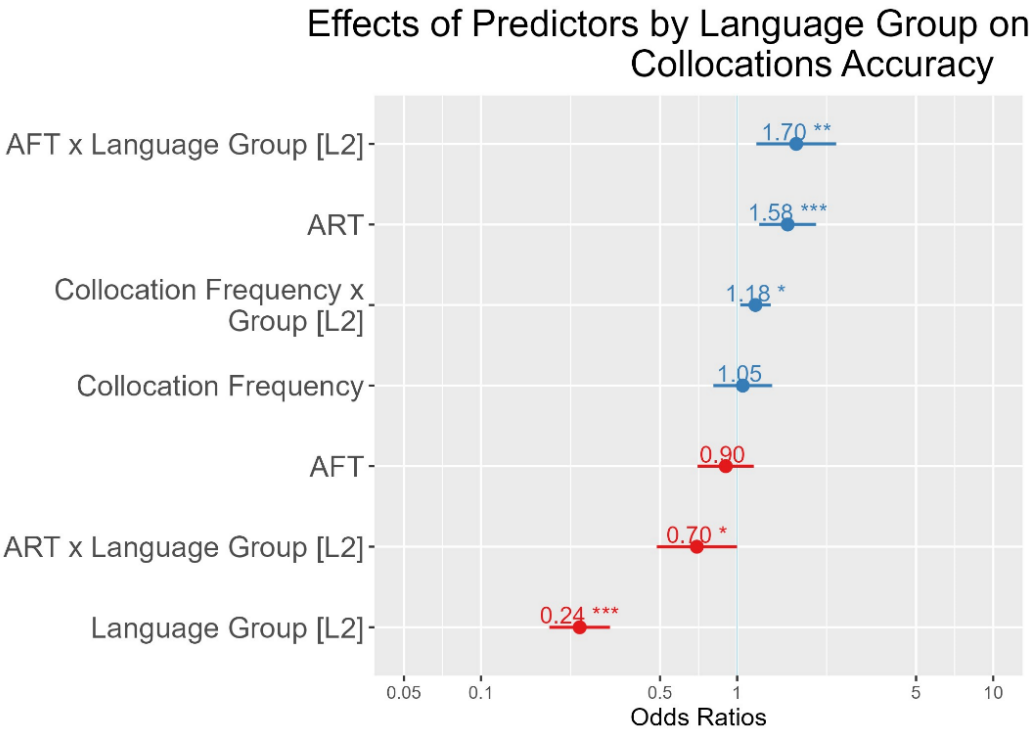


Figure 3.6: Fixed effects with interactions for language group on all variables, predicting odds of correct collocations selections. Red lines represent reduced odds of correct responses, blue lines represent increased odds.

3.4.3 Mediating effects of semantic fluency

To evaluate whether verbal fluency generally might explain our results, we re-recruited L2 participants for a test of semantic fluency with three different item categories: "animals", "grocery items", and "famous people". Some participants interpreted the instructions incorrectly, providing names of French supermarket chains instead of grocery items, and categories of public figures (e.g. actor, musician) instead of proper names, but we opted to keep these observations. We

removed one participant who entered all items in French. We compare both a combined measure with the sum of all scores, as well as the individual subtasks.

We first built a regression model where AFT, $F(1, 43) = 5.28$, $p < .05$, and the SF sum score, $F(1,43) = 31.14$, $p < .001$, co-predicted L2 connectives, $F(2, 43) = 18.21$, $p < .001$, Adj-R^2: .43. For L2 collocations, only AFT predicted the outcome, $F(1, 43) = 27.67$, $p < .001$, model Adj-R^2: .38, whereas SF was non-significant.

Analysis by subtask also revealed divergent outcomes, with a model predicting connectives with AFT and animal naming, $F(2, 43) = 21.54$, $p < .001$, Adj-R^2: .48, showing effects for AFT, $F(1, 43) = 29.85$, $p < .001$, and animals, $F(1, 43) = 13.23$, $p < .001$; another model comparing AFT and groceries, $F(2, 43) = 14.86$, $p < .001$, Adj-R^2: .38, showing effects for AFT, $F(1, 43)= 25.22$, $p < .001$, and groceries, $F(1, 43) = 4.50$, $p < .05$; and a model for AFT and public figures, $F(2, 43) = 12.17$, $p < .001$, Adj-R^2: .33, showed an effect for AFT, $F(1, 43) = 23.35$, $p < .001$), but a null effect for public figures. Analogous models predicting collocations from AFT and animals, groceries, and public figures only showed effects of AFT, which were broadly similar across categories, $F(1, 43) = 26.37$-28.00, all ps $< .001$.

3.5 General Discussion

We sought to determine if a semantic fluency task for author names in L2 (AFT) could serve as an alternative to ART as a measure of print exposure, and if reading experience in L1 and L2 were differentially associated with L2 vocabulary

knowledge. We hypothesised ART would predict collocations and AFT would predict connectives, reflecting the different kinds of memory required for each task. Evaluating the correct use of connectives requires not only word recognition, but also knowledge of their function. Conversely, evaluating collocations is a far more automatic process—either you know which words tend to co-occur more than others, or you do not. In fact, however, we found that AFT was a better predictor of both L2 connectives and collocations compared to ART.

The finding that AFT predicted additional variance for connectives scores beyond LexTALE (and was marginal for collocations), whereas ART did not, further underscores the importance of L2 reading for acquiring L2 vocabulary. Given the high variability in L2 print exposure and language ability, and the restrictive nature of ART, this isn't entirely surprising. That an open-ended measure like AFT performs well in this regard, however, even when accounting for L2 proficiency, is the primary contribution of the present research. Second language research is replete with discussions about how to access L2 learners' *"cultural capital"* (Bourdieu, 1986; Tunmer et al., 2006), yet when evaluating the role of print exposure in these populations, researchers haven't always acknowledged that the language experiences of L2 speakers rarely mirror those of English natives. Consequently, an effective and reliable proxy measure of L2 print exposure may not be the same as one used for L1. This is precisely what we demonstrate, with interactive models showing ART is most effective in L1, and AFT exceeding in L2.

Furthermore, measures of L2 proficiency and print exposure outperformed analogous L1 measures as predictors of L2 vocabulary. Logically, one's degree of exposure to a particular language should explain more about vocabulary knowledge

in that language, compared with exposure to another. But L1 experience is generally considered fundamental, laying the groundwork for learning additional languages (Sparks, 1995; Sparks et al., 2012). Again, our connectives measure is an adapted version of a task from a study in which L2 proficiency was predicted by an L1, but not L2 ART (Wetzel et al., 2020). We maintain that this was due to limited L2 exposure, which makes ART an unsuitable measure for beginner L2 speakers. Granted, L1 proficiency is likely a limiting factor for L2 novices, but the question of what distinguishes *advanced* L2 speakers is a separate one. Once the target language becomes relatively proficient, it follows that more extensive and naturalistic L2 exposure becomes critical.

Despite the criticisms of using ART in L2, as reviewed in the introduction, it still correlated with L2 vocabulary, though considerably less so than AFT. While ART may index L2 print exposure in advanced populations such as this one, its overlap with proficiency measures might lead researchers to infer null effects for print exposure when controlling for other tasks. However, we acknowledge that AFT is unlikely to be useful for novice learners either, given their limited L2 reading experience.

Additionally, we found that non-author semantic fluency mediated some, but not all variance explained by AFT. This varied by outcome measure and by subtask predictor. For connectives, only public figure naming was non-significant when controlling for AFT. Naturally, semantic fluency for authors and public figures both require recall of proper names, and we observe the expected outcome that author names are more informative than celebrity names, as the former index

reading experience whereas the latter reflect general cultural exposure. Conversely, no semantic fluency subtask predicted collocations when paired with AFT.

The divergent outcomes between models for the different vocabulary measures suggest that semantic fluency, or something associated with it, plays a larger role in the processing of connectives, and print exposure is more important for acquiring collocations. We suspect that if the variance explained by AFT in L2 were simply due to differences in semantic fluency, firstly, we would also observe some effect of semantic fluency for the collocations task. Secondly, a similar effect of AFT would likely also be seen in the L1 English population. But in fact, we see a sort of "inverted picture", in which ART is the better predictor for L1 speakers, and AFT outperforms in L2. It is possible that the role of semantic fluency is simply stronger in L2 than in L1, given the wider range of L2 skill generally, and recent research demonstrating L1 and L2 speakers are primarily differentiated by fluency rather than comprehension (Kuperman et al., 2023; Siegelman et al., 2024)—yet it is unclear why such variance would not have been sufficiently captured by LexTALE, which also measures lexical access. We contend the limiting factor for AFT is familiarity with authors (and consequently, serves as a proxy for reading experience) rather than semantic fluency generally, as indicated by the null effect of public figure naming when paired with AFT. Thus, we argue the Author Fluency Task reflects the additional level of engagement with reading necessary to become highly proficient in L2.

Before concluding, we note some limitations to our study. First, we calculated AFT scores using one point for each author, with no weighting for authors who are perceived to be more (or less) valuable to the reader. Perhaps more

popular author names are more likely to represent general cultural knowledge rather than personal reading experience—after all, as mentioned in the Introduction to this thesis, one need not have read any of Stephen King or Jane Austen's books for them to come readily to mind when thinking of authors, and they may be associated with Hollywood adaptations of their works rather than the original material. Developing weights for author names is a complex and delicate issue, but one which bears consideration.

Although LexTALE was a robust predictor of vocabulary knowledge, it may also have some limitations. LexTALE is a word recognition measure, and word knowledge is a multidimensional construct, with depth of word knowledge and meaning considered a better metric than knowledge or recognition of form (Jeon & Yamashita, 2022). As a lexical decision task, LexTALE only indexes knowledge of word form (and correspondingly, processing speed). Moreover, some evidence has suggested that although the LexTALE is a robust measure of vocabulary knowledge, it may not be reliable as a global proficiency measure in L2 (Puig-Mayenco et al., 2023). Thus, a more sensitive measure may be required to separate the effects of overall L2 proficiency, semantic fluency, and L2 print exposure.

By demonstrating the effectiveness of recall over recognition measures in L2 contexts, this study also reinforces previous findings which indicate that explicit recall tasks provide more accurate assessments of L2 language proficiency. As a dynamic and culturally adaptable measure, AFT has the potential to improve our understanding of how reading experience influences second language acquisition. Nevertheless, AFT will also require replication in diverse language populations, since our findings may be partially related to the close linguistic distance between

English and French. However, we suspect this is unlikely to completely explain the results, since the similarities between these two languages might instead lead to diminished effects of L2 reading experience. Similarly, it is possible our findings may only apply to English L2 speakers due to the global spread and influence of English, and future studies will determine if these findings generalise well to other target languages—naturally, we expect it would only be effective in languages with a similar culture of readership to English. Nevertheless, AFT allows second language learners to demonstrate their print knowledge without concern about social desirability bias on reading habits surveys, or guessing author names on ART.

3.6 Data Availability

Data and code used in analyses are available on OSF: https://osf.io/q62mt/

4. The Author Fluency Task as a measure of print exposure for L1 Korean speakers of English

Abstract

This chapter extends the findings of the previous by evaluating whether the Author Fluency Task (AFT) also serves as a proxy measure of print exposure in a population of speakers from a more distantly related language. L1 Korean speakers of English as L2 ($n = 62$) were recruited to carry out the same battery of tasks as the previous chapter, but with Korean equivalents of French tasks. Results showed that although the L2 AFT and ART both equivalently correlated with English collocations knowledge individually, only ART predicted additional variance beyond the L2 proficiency measure (LexTALE). AFT and ART both predicted connectives scores as well, and explained additional variance beyond LexTALE, yet neither print exposure measure was preferred. L1 Korean measures were non-significant predictors of L2 vocabulary knowledge, indicating little direct role for L1 print exposure or proficiency. We compare and contrast with the L1 French group and L1 English speakers from Chapter 3, as well as an additional sample of L1 English Oxford University students. Finally, we interpret the significance of these findings in light of the Korean educational context and linguistic distance accounts.

4.1 Introduction

In the previous chapter (McCarron et al., 2025), we demonstrated that in a sample of L1 French speakers of English as L2, the Author Fluency Task (AFT) was significantly more associated with L2 vocabulary knowledge than was ART. However, it is possible that these findings may have been partially due to the relative similarity in language and culture between English and French. Of course, much of the vocabulary of English is historically derived from French and/or Latin (approximately 50%; Gray & Atkinson, 2003)—to the extent that a recent book has controversially argued, firing the opening salvo with polemic flair from its very title, that " 'The English language does not exist': it is just mispronounced French" (« "*La langue anglaise n'existe pas": c'est du français mal prononcé* »; Cerquiglini, 2024).

Yet the kinds of formulaic language used in the previous study (i.e. discourse connectives and collocational word pairs) are less governed by formal rules than by convention, which naturally has evolved over the intervening centuries as the two languages have variously diverged and converged. Thus, as discussed in the previous chapter, an equivalent translation of any given connective may acquire subtle changes in meaning from one language to another. This may unintentionally alter the interpretation of a clause which follows it. As one example, Zufferey and Gygax (2017) describe the use of the French connective *« en effet »* (translated variously as "indeed", "in fact", or "because" in English), which can be used to express the coherence relations of both causality and confirmation in French. Because these relations are often mapped onto different connectives in other languages, acquiring *en effet* can prove challenging for French learners. Collocations

are similarly governed informally by custom and use, as word pairings are fundamentally arbitrary in the sense that many different complements might express an equivalent meaning, yet not all will sound correct to native speakers (recall the example of "raise prices" as opposed to "lift prices").

Despite this, it is clear that English and French share much in common, from vocabulary to writing system to culture. Because of this, it is not possible to generalise the findings of the previous chapter to other, more distantly related (or entirely unrelated) language populations. In light of this, the broader use of a print exposure measure such as AFT is contingent upon its validation in other groups. Having established its effectiveness in French natives, one approach might be to examine whether AFT is associated with language outcomes in L1 populations most unrelated to English. If successful, it would suggest reading for pleasure is critical for acquiring knowledge of English more generally, underscoring the importance of exposure outside of a formal classroom setting.

One potential candidate for an alternative language population to study is Korean speakers. South Koreans are generally very highly-educated, leading to a surge in economic growth in recent decades through increased access to global markets (Oh, 2010; J. C. Shin, 2012). This economic progress has also expanded interest in English learning and teaching, although this is not always easily accessible—for example, English courses in public schools are typically taught by non-natives, whereas private programmes can be prohibitively expensive (Magno, 2010). Moreover, the Korean alphabet (or, more accurately, "alphabetic syllabary", Pae, 2011; or "morphosyllabic alphabet", Pae et al., 2019), known as "Hangul" in South Korea, is a writing system completely unrelated to the Latin-Roman script

shared by French and English. Korean learners of English, then, have an additional barrier to entry in that they must not only acquire vocabulary and grammar from an unrelated spoken language, they must also learn a completely new orthography as well. This is a demanding task for any beginner, but compounding the difficulty, Korean speakers are accustomed to a highly transparent writing system, with phoneme-grapheme correspondences so predictable that it is possible to learn to read L2 Korean text as an adult in just a couple of hours (Pae, 2024b, 2024a). In contrast, English is somewhat notorious for its opaque orthography and confounding rule exceptions, which can be discouraging for learners (Cho, 2004).

From the reverse perspective of Korean as a second language, the ease of learning Hangul alone does not seem to translate into acquiring Korean as L2 more generally. Measures of linguistic distance which assess the relative ease of learning various languages as a native speaker of English place Korean firmly in the bottom position (1.00, shared with Japanese; even below Cantonese, 1.25; and Mandarin, 1.50), whereas French is situated near the top (2.50, compared to the maximum of 3.00 for Afrikaans, Norwegian, Romanian, and Swedish; Chiswick & Miller, 2005). This shared difficulty attests to the broad typological differences between the unrelated English and Korean languages. In contrast, English and French are merely close cousins descended from the Indo-European family of languages (Gray & Atkinson, 2003; Serva & Petroni, 2008).

Because of these differences, English reading may be less accessible for Korean compared to French natives, which makes them an interesting test case for the use of AFT as an indicator of L2 print exposure. Moreover, a Korean speaker's L1 proficiency may also be more associated with L2 English vocabulary outcomes

compared to French speakers, given that francophones are more likely to have opportunities for English language exposure outside of the classroom. Because of their geographic distance from majority English-speaking countries, not to mention the greater linguistic distance between the two languages, we anticipate that opportunities for English-language learning are more constrained to the classroom in South Korea, which in turn may be more accessible to those already more proficient in L1.

Despite the challenges of learning English, it has become essential for social mobility in South Korea (Choi, 2021; C. Lee, 2023). Accordingly, English proficiency tends to be relatively high, as shown on the EF English Proficiency Index which assesses the average language skill of English speakers across 113 countries globally (EF Education First, 2023). The most recent rankings showed South Korea (49) placing just a few spots behind France (43), both in the middle of the "moderate proficiency" range.[17] To give an example of what this means, the authors of the report suggest that an average English speaker in this proficiency band would be able to "participate in meetings in one's area of expertise", "understand song lyrics", and "write professional emails on familiar subjects" (p. 46). This level of L2 competence may be sufficient to serve its purpose in particular settings, even if speakers may generally lack a degree of nuance reserved for more

[17] However, much like the Test of English as a Foreign Language (TOEFL; ETS, 2024) scores, EF EPI rankings may be affected by factors such as self-selection bias (primarily younger people interested in language, with a median age of 25; EF Education First, 2023, p. 44) or how many respondents from a given country completed the test. In the case of the EF EPI, though, the company states they only include countries with at least 400 respondents, noting the actual figure would typically be much higher (EF Education First, 2023, p. 44).

advanced peers. Although South Korea placed lower than a few other Asian countries, namely Singapore (2), Philippines (20), Malaysia (25), and Hong Kong (China; 29), they placed considerably higher than many others, including (Mainland) China (82), Japan (87), and Thailand (101). Despite this, self-report ratings of English proficiency among the highest achieving Korean speakers may be considerably lower compared to their Test of English as a Foreign Language (TOEFL) scores (Im, 2018), reflecting a degree of self-doubt or uncertainty about their own language skills, potentially due to a lack of opportunities to use these skills socially.

Evidently, South Koreans tend to take education seriously, can become reasonably proficient in English, and are likely comparable to native French speakers on average. But are individual differences in print exposure, or "free voluntary reading" a significant contributor for learning English in South Korea, or is second language learning primarily restricted to a classroom setting? Looking at this question, Kim and Krashen (1998) examined a group of Korean students who spoke English as L2, and showed that scores on the English ART were significantly associated with vocabulary skill. They found that ART explained 38% of the variance on test scores, whereas including scores from a Magazine Recognition Test (MRT) and a self-report measure of reading in a hierarchical regression model explained just an additional 6%, suggesting ART on its own was most useful in practice. Perhaps contrary to the assertions of McCarron and Kuperman (2021) then, Kim and Krashen's findings would indicate that ART may still be useful in certain targeted L2 populations with significant English exposure. Rather, it may be that ART is most unreliable when used in a diverse cohort of language

backgrounds, as aggregating them into a singular L2 group may hide meaningful differences between them.

Some research has also looked at the role of print exposure in relation to L1 Korean proficiency. Inspired by work on the Chinese Author Recognition Test (CART; Chen & Fang, 2015), Lee and colleagues (2019) devised a Korean version (KART). Scores correlated with Korean vocabulary and comprehension knowledge, and they predicted a decreased word familiarity effect in a Korean lexical decision task, with higher KART scores associated with greater accuracy in the unfamiliar word condition. The authors also evaluated the role of primary vs. secondary print knowledge (Martin-Chang & Gould, 2008), asking participants to indicate authors they had selected on the KART who they had read personally. Noting that very few of the authors recognised on the KART had been read personally by participants, Lee and colleagues (2019) suggested that Korean students likely read very few books in L1, and instead tend to learn about authors through secondary exposure in the classroom. If this is the case, it is likely also true of Korean speakers' English print exposure, and this may reflect an important difference between how Korean and French speakers learn English as a second language. With these considerations in mind, we introduce the present study.

4.1.1 Present study

The present study received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee [reference R77364/RE003] and was pre-registered on OSF (https://osf.io/zgh8f). We tested

whether L1 Korean / L2 English reading experience (as assessed by AFT and ART in both languages) is associated with individual differences in English vocabulary knowledge, even when controlling for proficiency. Our research questions were broadly intended to assess the utility of AFT as a measure of print exposure in a more distantly related language population. These questions were:

1) Are scores on an L2 English AFT more highly correlated with English vocabulary knowledge than ART? Does either measure explain additional variance not accounted for by a general proficiency measure (LexTALE)?
2) Does L1 Korean or L2 English print exposure better predict performance on English vocabulary tasks?
3) Do AFT/ART perform differently according to the vocabulary outcome measure (collocations vs. connectives)?

For our L1 Korean / L2 English cohort, we made the following predictions, which we have separated by L1 and L2 predictors and simplified for interpretability:

L2 Predictors

H1a: LexTALE scores would positively predict connectives scores.

H1b: LexTALE scores would positively predict collocations scores.

H2a: AFT scores would positively predict connectives scores.

H2b: AFT scores would remain significant as a positive predictor of connectives scores when controlling for LexTALE, but ART would not.

H3a: ART scores would positively predict collocations scores.

H3b: ART scores would remain significant as a positive predictor of collocations scores when controlling for LexTALE, but AFT would not.

L1 Predictors

H4a: Korean LexTALE scores would positively predict English connectives scores.

H4b: Korean LexTALE scores would positively predict English collocations scores.

H5a: Korean ART scores would positively predict collocations scores.

H5b: Korean ART scores would remain significant as a positive predictor of collocations scores when controlling for the Korean LexTALE.

Essentially, we suspected that:

a) Korean/English LexTALE scores would both positively predict English connectives and collocations scores.
b) English AFT (but not ART) scores would positively predict connectives scores when controlling for the English LexTALE; and
c) Korean/English ART (but not AFT) scores would both positively predict collocations scores when controlling for LexTALE in either language.

The rationale for a) was that proficiency in both languages might contribute to both L2 outcome measures, with L1 proficiency indirectly supporting second

language learning. For b), we reasoned that the L2 AFT, as an explicit memory/recall measure of print exposure, would be more associated with the connectives task, which would require more careful analysis to determine which item best fit into the sentence, whereas for c), we reasoned that the collocations task would rely on the more passive recognition memory shared with ART. Additionally, we did not anticipate that either measure of Korean print exposure would predict English connectives.

In this chapter, we contextualise the performance of our newly recruited Korean native speakers (often referenced in shorthand as "KR") with reference to the L1 French data ("FR") and our original sample of L1 English speakers from the general population of the United Kingdom ("EN") from Chapter 3. In addition, we include a sample of L1 UK English speaking Oxford University students ("EN-OX"). This approach enables us to contrast two populations who likely acquire English as L2 through different learning strategies, as well as how L1 Korean speakers compare to both general and academic L1 English populations.

4.2 Methods

4.2.1 Participants

L1 Korean/L2 English participants (n = 62, 40 women) were recruited online via Prolific (2024). Upon providing informed consent, they were directed to complete the study on Gorilla (Anwyl-Irvine et al., 2020). Participants were reimbursed for their time at a rate of £10 per hour. All data was anonymised. Pre-screening criteria

required participants to be between 18-75 years of age, with normal or corrected-to-normal vision, who had not taken part in any prior pilot studies. Our complete pool of participants ranged from 18-68 years of age (M = 30.52, SD = 9.21). For comparison, the sample of Oxford University L1 English students (n = 36) which we refer to throughout this chapter were between 18-23 years old (M = 19.06, SD = 0.98).

4.2.2 Measures

English versions of AFT, ART, LexTALE, connectives, collocations, and motivation tasks used in this study were the same as described in the previous chapters (see section 2.2.2). Below, we detail only the Korean versions of tasks where appropriate.

AUTHOR FLUENCY TASK (KOREAN)

The Korean AFT was identical in procedure and scoring to the English version described in Chapter 3 (McCarron et al., 2025), but with Korean-language instructions. Accordingly, participants were asked to provide names of authors which have been published in Korean, either originally or in translation.

AUTHOR RECOGNITION TEST (KOREAN)

The Korean ART follows the same procedure and scoring logic as the English version described in Chapter 3, but participants were provided instructions in Korean. This version was taken from Lee et al. (2019), and features 40 author names and 40 foil names (see Appendix A.8). Of the 40 authors included, 22 were determined to be "foreign" authors, or those who originally wrote in a language other than Korean. Although Lee and colleagues refer to their test as the "KART", we use "ART-KR" to remain consistent with the French ART (ART-FR).

LEXTALE (KOREAN)

The Korean LexTALE (Son et al., 2022) follows the same procedure and scoring logic as the English version described in Chapter 3, but contains 57 Korean words and 41 plausible Korean non-words (98 items total, see Appendix A.11). Instructions were provided in Korean, and performance was assessed as the percentage of correct selections out of the total.

4.2.3 Procedure

Participants began by completing the 1) demographics questionnaire, followed by 2) the English and Korean AFT, 3) the English and Korean LexTALEs and Author Recognition Tests, 4) the connectives and collocations tasks, and 5) the motivation survey. Task order was counterbalanced for levels 2, 3, and 4 due to task

similarities. The L1 participant procedure was identical, excluding Korean tasks. The order of tasks is illustrated in Figure 4.1.

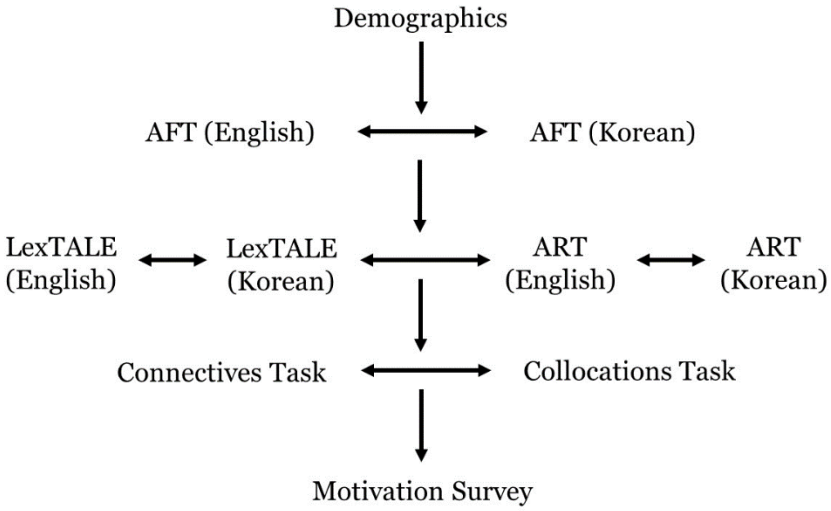


Figure 4.1: Order of tasks for the L1 Korean cohort. Tasks which were counterbalanced are indicated by horizontal lines. The L1 English participant procedure was identical, excluding Korean tasks.

4.3 Results

4.3.1 Demographics

The bar plots below highlight key demographic findings. In Figure 4.2, we observe that most participants rated their English proficiency very highly, with 22 stating they possessed a level of "Full Professional Proficiency" and an equal number at "Native / Bilingual". Figure 4.3 shows respondents indicated a generally high appreciation for the role of reading as a contributor to developing their English

proficiency, and Figure 4.4 shows most participants began learning English before the age of 10.

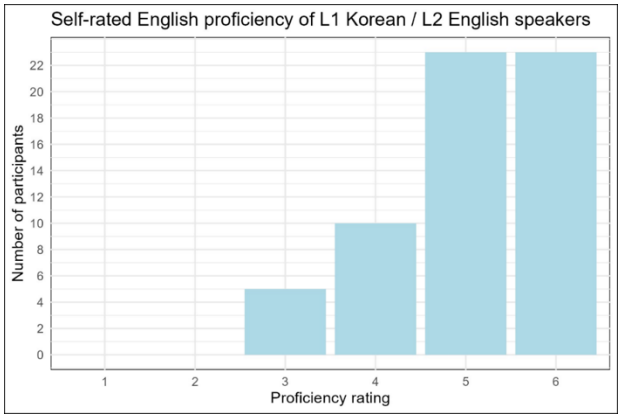


Figure 4.2: Subjective English proficiency ratings for L1 Korean speakers of English (1 – No Proficiency or Not Applicable; 2 – Elementary Proficiency; 3 – Limited Working Proficiency; 4 – Professional Working Proficiency; 5 – Full Professional Proficiency; 6 – Native / Bilingual Proficiency).

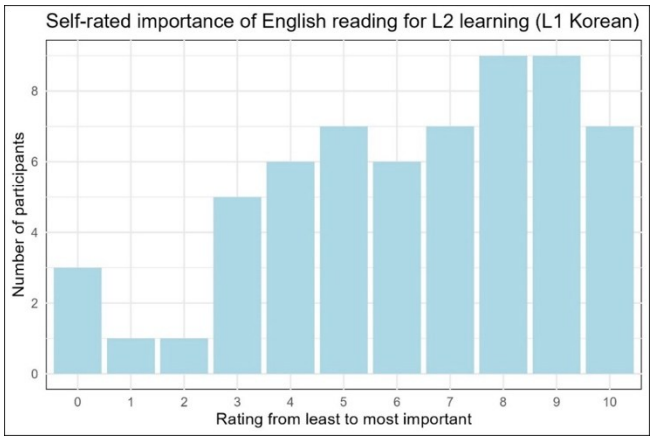


Figure 4.3: Subjective ratings of perceived importance of L2 English reading for L2 English learning, L1 Korean participants.

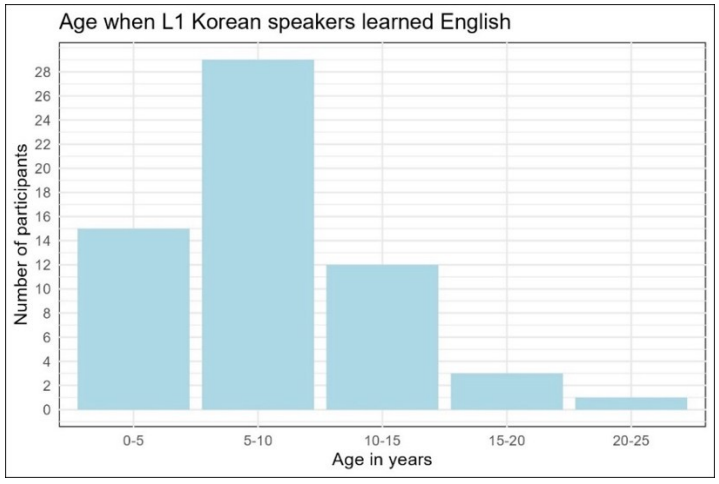


Figure 4.4: Age in years at which L1 Korean speakers learned English.

4.3.2 Descriptive Statistics

As in the previous study, extreme outliers were identified as those falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR within their respective cohort. This was not preregistered, but we determined it was necessary due to low scores for certain tasks, and to enable consistent comparisons between language groups. Outlier observations were subsequently removed at the task level, leading to slightly lower sample sizes in some measures. For consistency, we applied the same rule for Korean participants, which led to a reduction in sample size; n was lowest in the Korean LexTALE ($n = 52$), where performance was relatively poor.

Table 4.3.1 lists summary statistics for each test by language group. Comparing the different groups, it is not surprising that L1 participant scores consistently exceeded L2, most notably in the LexTALE and collocations tasks.

Table 4.3.1: Summary statistics for each test by language group. Because some measures failed to meet assumptions of normality and homogeneity of variances, we compare performance for each using the Mann-Whitney U (or Wilcoxon rank-sum) test. Each U test compares the row with the L1 Korean group (KR). To correct for multiple comparisons, *p*-values were Bonferroni corrected.

	Group	n	Min	Med.	Mean	Max	SD	IQR	*U*	*p*-value
AFT (EN)	EN	59	3	12	12.56	27	5.49	8	2709.50	< **.001*** **
	EN-OX	36	6	14	14.89	29	5.31	7.5	1867.00	< **.001*** **
	FR	60	1	8.50	8.92	22	5.32	9	2060.00	1
	KR	62	-1	7	7.9	23	5.06	8	-	-
AFT (KR)	KR	59	-1	4	4.9	14	3.81	6	-	-
ART (EN)	EN	60	5	25.50	27.23	53	11.18	14.75	2802.50	< **.001*** **
	EN-OX	36	15	30.50	29.81	49	10.02	14.25	1815.50	< **.001*** **
	FR	60	0	20	21.35	46	10.40	15.25	2295.50	.46
	KR	62	1	16	17.31	45	9.99	12.5	-	-
ART (KR)	KR	62	-2	15	14.94	33	9.43	12.75	-	-
Collocations	EN	59	42.50	77.50	75.59	92.50	11.40	20.00	3263.00	< **.001*** **
	EN-OX	35	60.00	70.00	70.00	82.50	6.94	10.00	1855.50	< **.001*** **
	FR	60	20.00	50.00	49.33	85.00	17.55	33.12	1789.00	1
	KR	62	12.50	52.50	50.65	82.50	16.93	23.75	-	-
Connectives	EN	59	61.67	86.67	85.28	98.33	8.68	13.33	2218.50	**.03***
	EN-OX	35	78.33	90	89.90	100.00	5.10	5.83	1577.50	<**.001*** **
	FR	60	20.00	79.17	71.97	100.00	21.77	35.42	1516.50	1
	KR	56	58.33	80.00	79.73	98.33	9.46	11.67	-	-
LexTALE (EN)	EN	53	88.33	96.67	95.88	100.00	3.25	5.00	2735.00	< **.001*** **
	EN-OX	34	88.33	95.00	95.05	100.00	3.16	3.33	1693.00	< **.001*** **
	FR	60	60.00	84.17	83.11	100.00	11.25	20.00	1654.00	1
	KR	61	60.00	86.67	84.95	100.00	10.95	13.33	-	-
LexTALE (KR)	KR	57	77.55	95.92	94.75	100.00	5.48	6.12	-	-
Motivation	EN	59	3.30	4.30	4.27	5.00	0.40	0.55	2979.00	< **.001*** **
	EN-OX	32	2.70	3.80	3.76	4.60	0.53	0.80	1060.50	1
	FR	59	3.30	4.10	4.11	4.90	0.34	0.45	2749.00	< **.001*** **
	KR	62	2.70	3.75	3.71	4.80	0.48	0.70	-	-

Figure 4.5 illustrates the density distributions for the different measures used in the Korean group in this study. All measures are shown in their original scales, and generally reveal non-normal distributions.

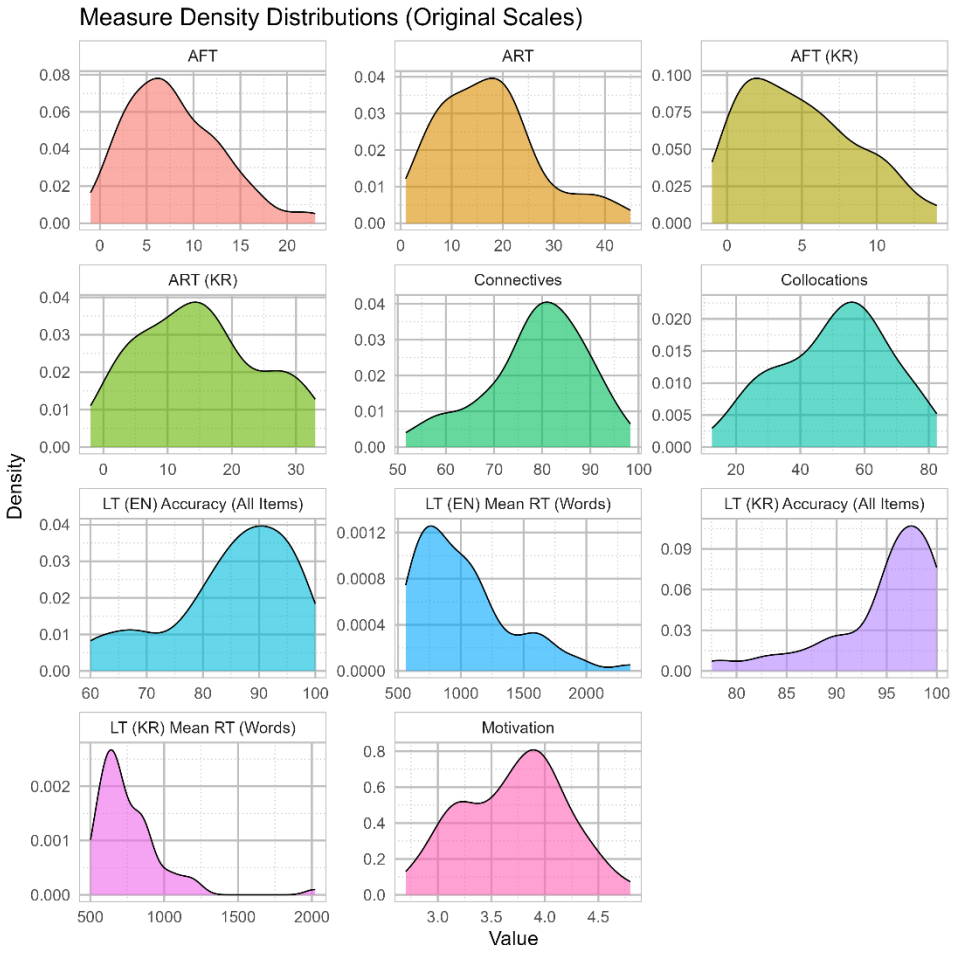


Figure 4.5: Density plots of each measure in their original scales, L1 Korean sample.

PRINT EXPOSURE MEASURES

Author Fluency Task

Figure 4.6 illustrates performance on AFT across language groups as raincloud plots. As the plots suggest, and the pairwise comparisons confirm (Table 4.3.1), the

two L2 groups did not significantly differ on their raw AFT scores (L1 Korean M = 7.9, SD = 5.06; L1 French M = 8.92, SD = 5.32). However, both L1 English groups performed significantly better than the L2 groups.

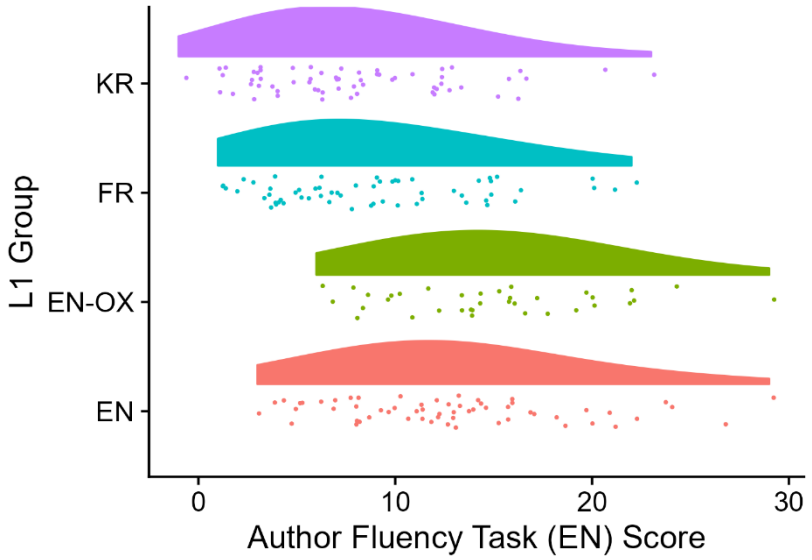


Figure 4.6: Raincloud plots showing distributions of observations for AFT scores by L1 group.

Author Fluency Task (Korean)

L1 Korean participants performed more poorly (M = 4.9, SD = 3.81) on their native language AFT compared to the L2 task. A list of the 60 most commonly named authors on the Korean AFT can be found in Appendix B.2.5.[18]

Author Recognition Test

Reliability for ART (KR-20) was excellent in all cohorts (KR: 0.94, SEm: 2.42; FR: 0.94, SEm: 2.55; EN: 0.94, SEm: 2.62; EN-OX: 0.93, SEm: 2.63). L1 Korean participants performed modestly on this test, but with wide dispersion (M = 17.31, SD = 9.99; see Table 4.3.1). As with AFT, the L2 English groups did not perform significantly differently on ART scores, but still trailed behind English natives (illustrated in the raincloud plot in Figure 4.7).

[18] Interestingly, while cleaning the typed data on the Korean AFT, we noted that no spelling corrections were required, likely due to the transparent orthography of the Korean writing system. For comparison, regardless of the language group under examination, the English and French versions of this task required many hundreds of lines of code and manual intervention to correct spelling errors and ensure the names were consistent.

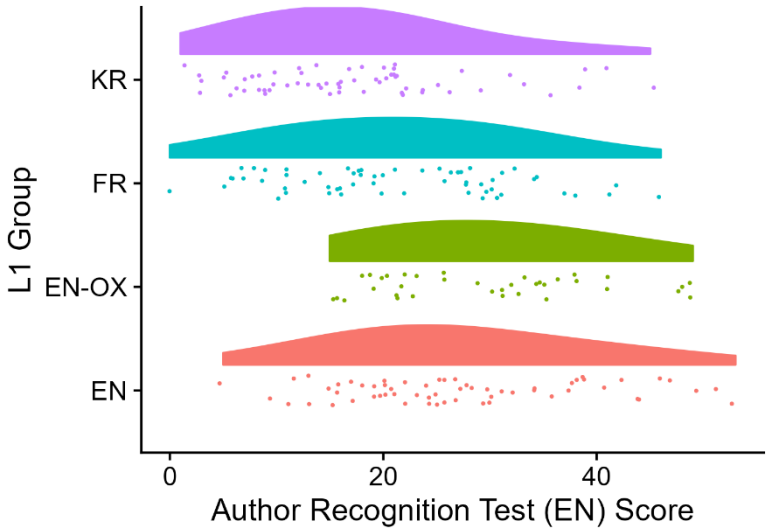


Figure 4.7: Raincloud plots showing distributions of observations for ART scores by L1 group.

A full list of English ART name selections by L1 Korean participants, along with response time statistics can be found in Appendix B.1.3, and Figure 2.2 illustrates the total number of selections. Korean participants recognised just 14 out of 60 names (23%) at a rate of 50% or more, indicated by the red line in this figure. For comparison, the rate for L1 French speakers was 21/60 (35%), L1 UK English speakers 28/60 (46.66%), and L1 English Oxford participants had a rate of 32/60 (53.33%).

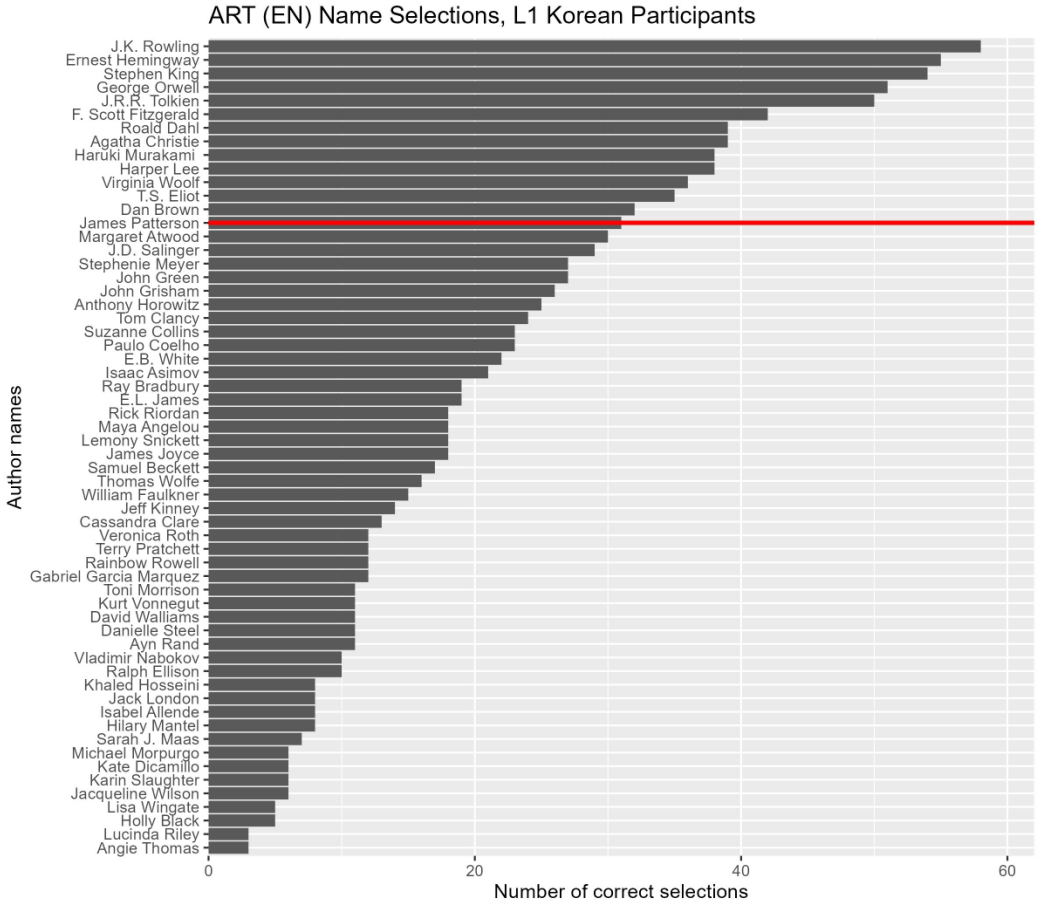


Figure 4.8: Number of author name selections on the English ART, L1 Korean participants (total $n = 62$). Names at or above the red line (James Patterson) are those with a recognition rate greater than 50%.

Author Recognition Test (Korean)

Reliability for the Korean ART was excellent (KR-20: 0.97, SEm: 1.72). L1 Korean participants performed modestly on this test, but with wide dispersion ($M = 14.94$, $SD = 9.43$; see Table 4.3.1), and performed slightly lower on average compared to the original paper by Lee et al. (2019), where the average number of authors recognised was 17 ($SD = 6.7$). In our study, participants recognised just 14/40

(35%) of author names at a rate of 50% or higher (Figure 4.9), although this is higher than the corresponding rate for the L1 ART (23%). A full list of Korean ART name selections, along with English translations and response time statistics can be found in Appendix B.1.5.

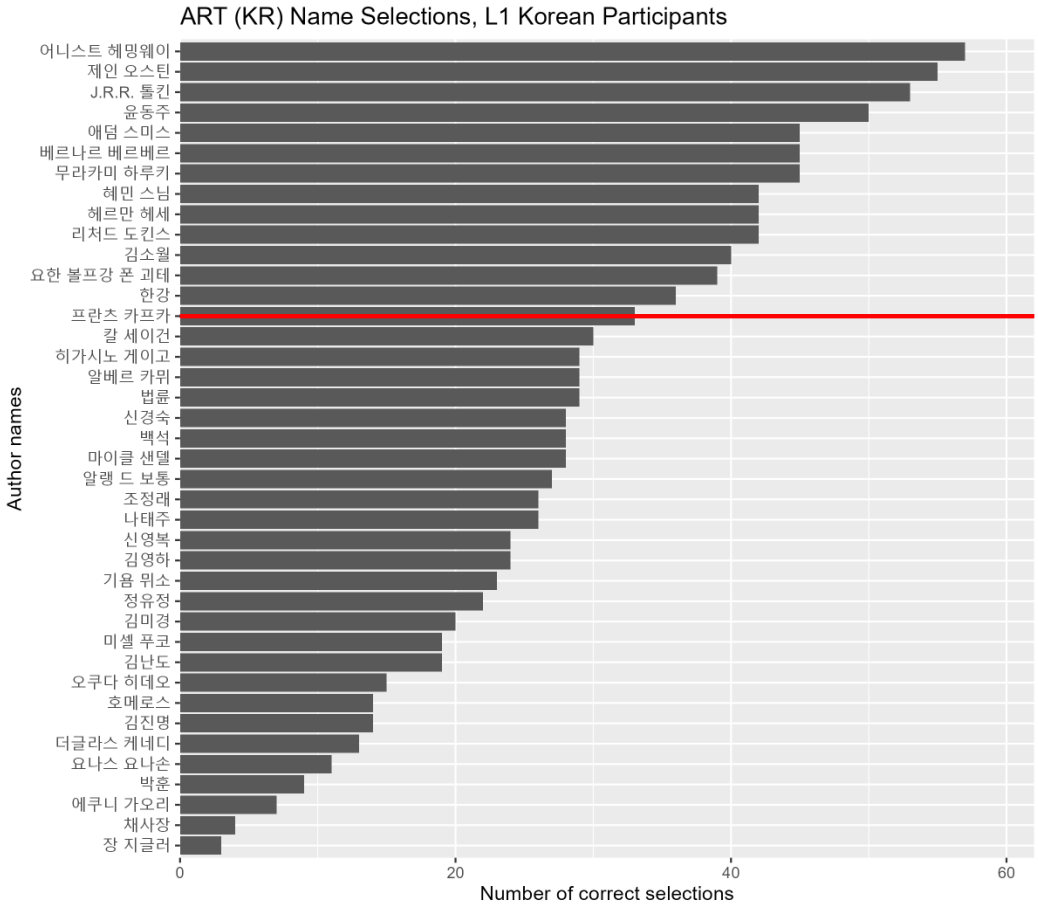


Figure 4.9: Number of author name selections on the Korean ART, L1 Korean participants (total $n = 62$). Names at or above the red line (프란츠 카프카, "Franz Kafka") are those recognised by fewer than 50% of participants.

VOCABULARY MEASURES

LexTALE

In the L1 Korean sample, reliability (KR-20) for the English LexTALE was high (0.86, SEm: 1.94). For comparison, this was similar to the L2 French sample (0.86, SEm: 2.10), whereas reliability was acceptable in L1 English general population sample (0.76, SEm: 1.23), and was poor in the L1 English Oxford student sample (0.58, SEm: 1.67). This comparatively lower reliability in L1 samples likely reflects the fact that LexTALE is primarily intended to assess second language speakers.

L1 Korean participants scored relatively high ($M = 84.95$, $SD = 10.95$), and were equivalent to L1 French speakers ($M = 83.11$, $SD = 11.25$), though both groups performed below native speakers of English on average (see Table 4.3.1). The raincloud plots in Figure 4.10 illustrate the performance ranges and distributions of each language group. A full list of the LexTALE stimuli and performance (percentage accuracy and mean response times) is provided in Appendix B.4.1.

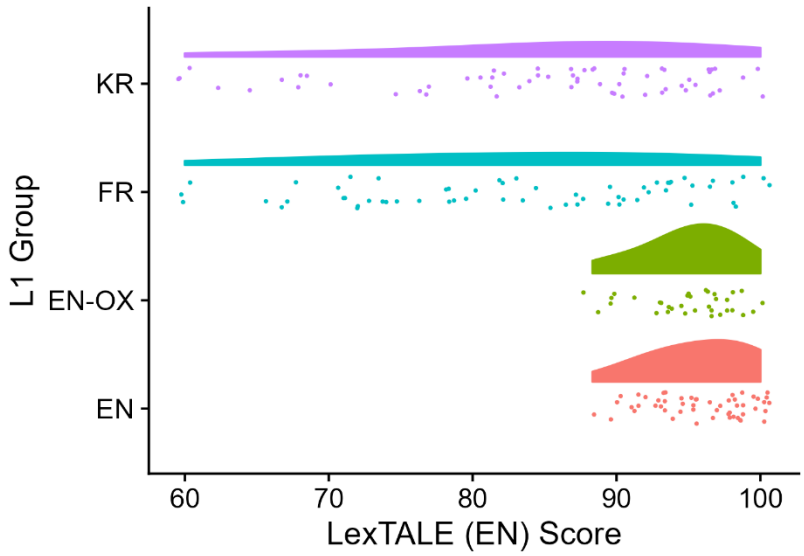


Figure 4.10: Raincloud plots showing distributions of observations for English LexTALE scores by L1 group.

LexTALE (Korean)

Participants fared very well on the Korean LexTALE ($M = 94.75$, $SD = 5.48$), and scores were generally higher than the equivalent L2 English measure. Reliability was also excellent (KR-20: 0.94, SEm: 1.80). A table with statistics for all Korean LexTALE items can be found in Appendix B.4.3.

Connectives

Reliability (KR-20) for this task was good to excellent in both L2 groups (KR: 0.89, SEm: 2.85; FR: 0.95, SEm: 2.94), though it was somewhat lower in L1 English samples (EN: 0.84, SEm: 2.47; EN-OX: 0.75, SEm: 2.16). Average scores for each connective, by coherence relation, frequency, and L1 group are presented in

Table 4.3.2. Compared to the L1 French group (M = 71.97, SD = 21.77), the L1 Korean group generally scored higher on this task, and with much lower spread (M = 79.73, SD = 9.46), although this was not significant after the Bonferroni correction (see Table 4.3.1). Once again, L1 speakers predictably outperformed second language users. Distributions of scores across all connectives by language group are shown as raincloud plots in Figure 4.11, and percentage of correct selections within each language group by coherence relation are illustrated in Figure 4.12.

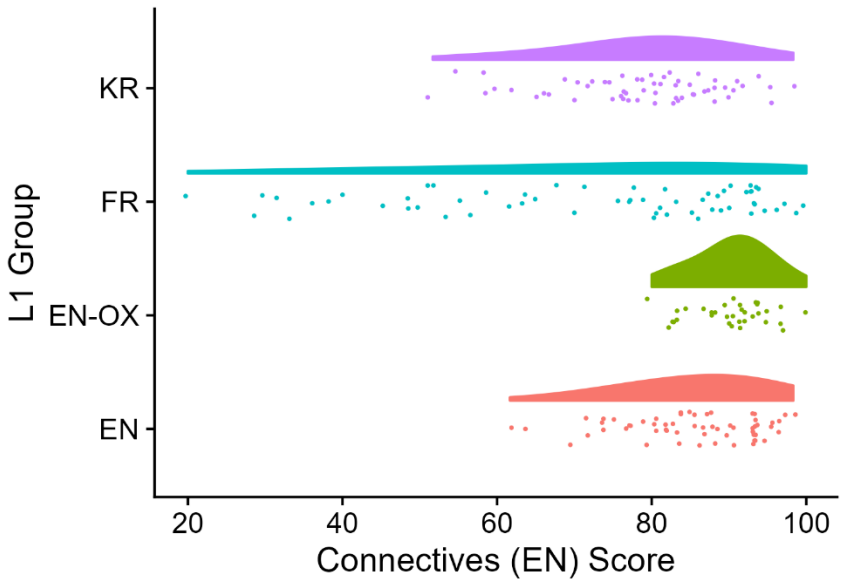


Figure 4.11: Raincloud plots showing distributions of observations for connectives scores by L1 group.

Table 4.3.2: Accuracy scores as percentages per connective, by frequency (high/low) and separated by L1 group.

Relation	Item	Freq	EN		EN-OX		FR		KR	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Addition	indeed	high	0.52	0.50	0.61	0.49	0.58	0.49	0.44	0.50
	furthermore	low	0.83	0.37	0.93	0.26	0.76	0.43	0.85	0.35
Cause	since	high	0.91	0.29	0.94	0.23	0.75	0.43	0.89	0.31
	given that	low	0.94	0.25	0.95	0.22	0.71	0.45	0.75	0.43
Concession	despite this	high	0.87	0.34	0.95	0.22	0.8	0.4	0.86	0.34
	nevertheless	low	0.90	0.30	0.92	0.27	0.8	0.4	0.89	0.32
Condition	as long as	high	0.93	0.25	0.86	0.35	0.83	0.38	0.83	0.38
	provided	low	0.90	0.30	0.84	0.37	0.61	0.49	0.56	0.50
Consequence	therefore	high	0.90	0.30	0.97	0.17	0.79	0.41	0.90	0.30
	hence	low	0.79	0.41	0.97	0.18	0.62	0.49	0.85	0.36
Contrast	whereas	high	0.91	0.29	0.95	0.21	0.77	0.42	0.83	0.38
	conversely	low	0.85	0.36	0.91	0.29	0.62	0.49	0.91	0.28
Global	***Overall***		**0.85**	**0.35**	**0.91**	**0.29**	**0.72**	**0.45**	**0.80**	**0.40**
	High Frequency		**0.84**	**0.37**	**0.89**	**0.32**	**0.75**	**0.43**	**0.79**	**0.41**
	Low Frequency		**0.87**	**0.34**	**0.92**	**0.26**	**0.69**	**0.46**	**0.80**	**0.40**

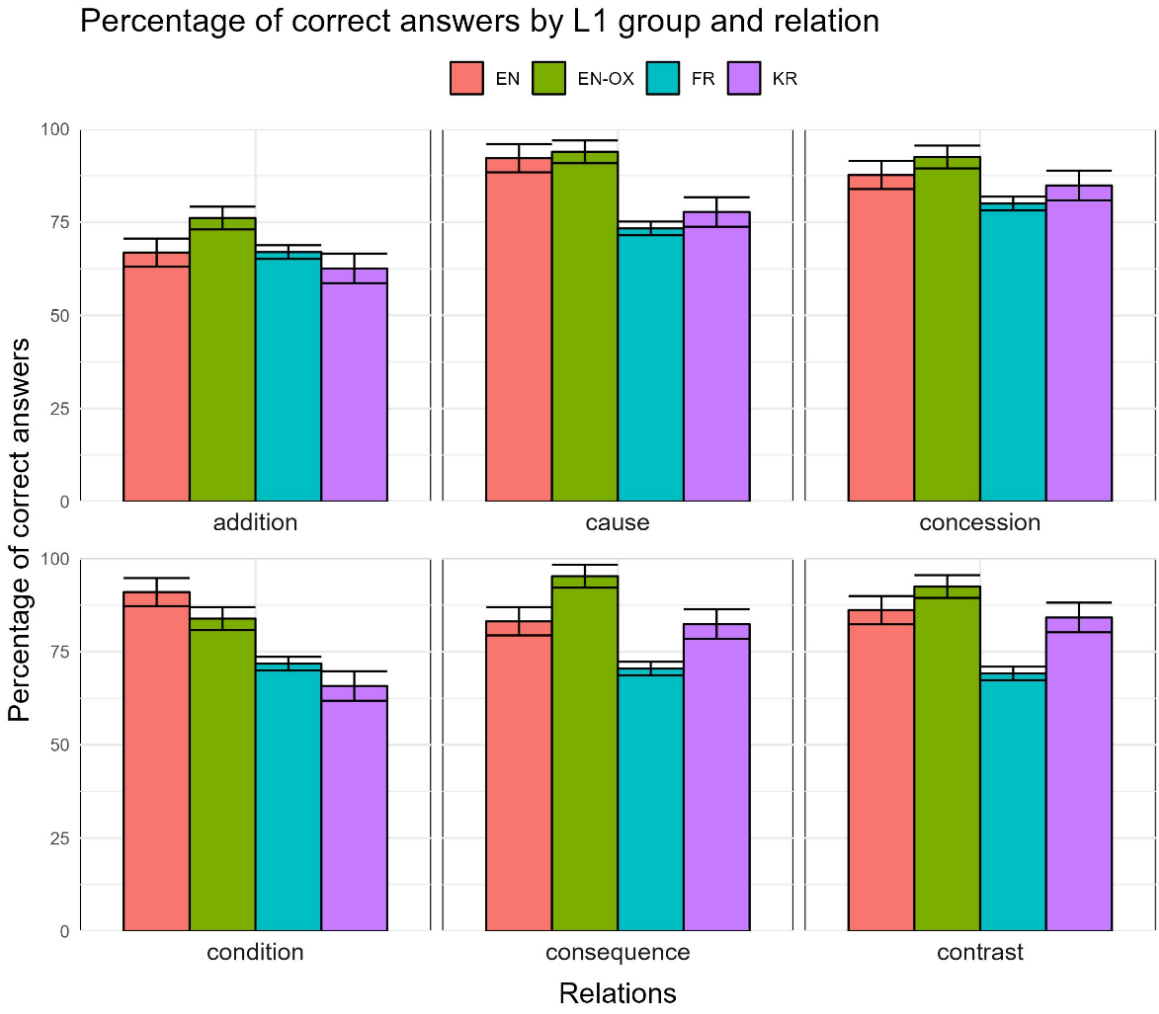


Figure 4.12: Percentage of correct answers per each coherence relation by language group. Error bars represent the standard error.

Intriguingly, we observe that Korean and French L2 speakers of English vary in terms of their performance on some coherence relations, with the Korean group generally performing higher than the French natives, most notably for relations of "consequence" and "contrast", but generally performing lower on "addition" and "condition".

Similarly, our L1 English groups showed different strengths. Compared to the sample of UK English speakers, Oxford undergraduate students performed higher on the "addition" relation, with significantly better scores even for the more challenging "indeed" connective. This observation may reflect its use as a more academic type of connective—accordingly, Oxford undergraduates likewise outperform on the relations of "consequence" ("therefore", "hence"), "contrast" ("whereas", "conversely"), and "concession" ("despite this", "nevertheless"), scoring virtually at ceiling (M = 0.93-0.97, SD = 0.17-0.25). On the other hand, the general UK population sample performed comparatively better on conditional connectives such as "as long as" and "provided".

Collocations

Compared to the connectives task, collocations scores were markedly lower across all L1 and L2 samples. L2 groups performed equivalently on the collocations task, with both scoring approximately 50% on average (see Table 4.3.1). Figure 4.13 illustrates performance on the collocations task as raincloud plots by language group.

The L1 UK English general population sample performed above the other groups, including the Oxford sample. This contrasts with performance on the connectives task, where the university students performed somewhat higher than the general UK population. As discussed in the previous chapter, this likely relates to how connectives and collocations are learned, with collocations naturally requiring a significant amount of additional exposure. Accordingly, we suspect this

reflects increased lifetime exposure to English in this comparatively older population (M_{age} = 39.42, SD = 14.39) compared to the Oxford students (M_{age} = 19.06, SD = 0.98). However, although reliability (KR-20) for this task was good in the L1 Korean (0.82, SEm: 2.81), L1 French (0.84, SEm: 2.74), and L1 UK English sample (0.83, SEm: 2.43), in the L1 Oxford sample reliability was extremely low (0.32, SEm: 2.57), perhaps due to the smaller sample size, as well as the generally shorter period of lifetime exposure to collocations in the younger sample.

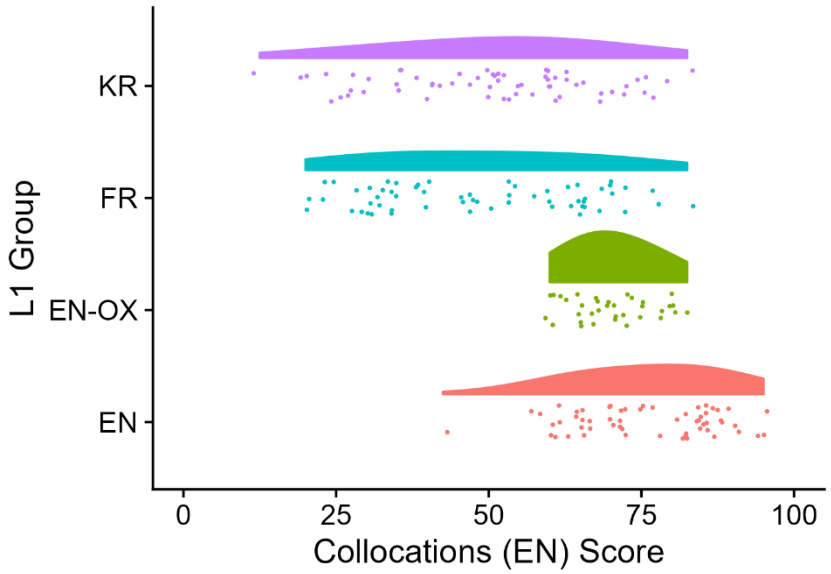


Figure 4.13: Raincloud plots showing distributions of observations for collocations scores by language group.

The full list of collocation selections by language group, along with item characteristics, including frequency and mutual information indices from Dąbrowska (2014) can be found in Appendix B.5.

4.3.3 Statistical Analysis

Below, we begin by showing Spearman correlations between all measures of interest (Table 4.3.3) before describing individual statistical models for each of our vocabulary outcome measures.

Table 4.3.3: Spearman correlation (ρ) matrix for all predictor and outcome variables of interest, L1 Korean cohort.

	AFT	*ART*	*AFT -KR*	*ART -KR*	*Coll.*	*Conn.*	*LT*	*LT -KR*	*Mot.*
AFT	1								
ART	**.582 *****	1							
AFT-KR	.252	.170	1						
ART-KR	.228	.023	**.473 *****	1					
Collocations	**.350 ****	**.350 ****	-.145	-.224	1				
Connectives	**.476 *****	**.499 *****	-.130	.013	**.391 ****	1			
LT	**.388 ****	**.344 ****	-.210	-.223	**.631 *****	**.432 *****	1		
LT-KR	-.078	-.223	.066	**.487 *****	-.209	-.185	-.239	1	
Motivation	-.014	.138	-.142	-.042	**.282 ***	.141	.150	.006	1

LT = LexTALE accuracy score for all stimuli. Significant correlations are in **bold**; * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

As Table 4.3.3 indicates, the Korean language independent variables are only correlated with each other, with no relationship to our English dependent variables. As a consequence, we find evidence for the null for both H4 and H5, which anticipated an effect of LexTALE-KR and ART-KR on the outcome variables, and we only report on models with English language predictors below. In contrast, we note that both the English AFT and ART are approximately equally correlated with connectives, collocations, and LexTALE scores.

CONNECTIVES

We began by building multiple regression models which evaluated the effects of LexTALE, AFT, and ART, which were all similarly and strongly correlated with connectives scores (Table 4.3.3). Our relevant hypotheses were that connectives scores would be predicted by LexTALE (H1a), AFT (H2a), and that the effect of AFT would remain when controlling for LexTALE, whereas ART would not (H2b).

To account for the different scales used, we first standardised our predictors. Our best model (Adj-R^2 = .29, $p < .001$; Table 4.3.4) showed an effect of LexTALE ($\beta = 3.24$, $p < .05$) and AFT ($\beta = 3.93$, $p < .01$), confirming H1a and H2a respectively, and partially confirming H2b which stated AFT would remain significant when controlling for LexTALE. However, estimates for an analogous model with AFT replaced with ART were virtually identical, and a Vuong test indicated no preference for either model (z = -0.181, p-values H1A = .57, H1B = .43), which contradicted H2b, as it was also hypothesised that the effect of ART would no longer be significant when controlling for LexTALE. Including both print

exposure measures led to non-significant main effects for both. To illustrate the individual effects of each predictor, standardised partial effects for a full model including LexTALE, AFT, and ART are illustrated in Figure 4.14.

Table 4.3.4: Multiple regression model predicting connectives scores by print exposure measures (L1 Korean participants).

Predictors	*Estimates*	*CI*	*p*
(Intercept)	78.822	76.489 – 81.155	**<0.001*****
LexTALE	3.242	0.751 – 5.732	**0.012***
AFT	3.930	1.439 – 6.421	**0.003****
Observations		58	
R^2 / R^2 adjusted		0.311 / 0.286	

** p<0.05 ** p<0.01 *** p<0.001*

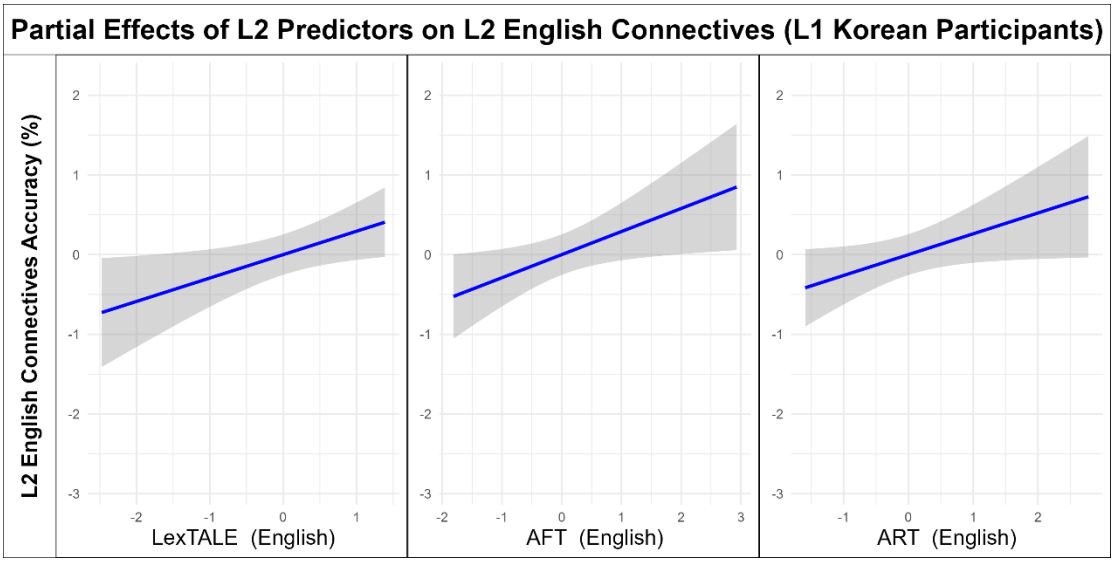


Figure 4.14: Standardised partial residual plots of all predictors on English connectives scores, L1 Korean participants. Shaded areas represent the 95% confidence interval.

In light of the findings of Chapter 3, we considered that the two print exposures may be complementary, or measure different aspects of reading experience. To examine this question, we created an exploratory model with an interaction for ART on AFT. This model showed improved fit (Adj-R^2 = .33, $p < .001$), with significant main effects for both LexTALE ($\beta = 2.68$, $p < .05$), and AFT ($\beta = 3.65$, $p < .05$), although the main effect of ART ($\beta = 2.75$, $p = .10$) and its interaction with AFT (β = -1.82, $p = .06$) were only marginal. Despite this, we consider that with a larger sample size, the negative interaction could suggest that the two measures reflect different kinds of print exposure. In other words, as the ART score increases, the effect of the AFT score might decrease, and vice-versa. This could

imply, for example, that those who provide more names on AFT may not be as familiar with authors on ART, underscoring the psychometric differences between the two measures which require different forms of memory.

Next, we built exploratory generalised linear mixed effects models (GLMER) to determine the influence of random effects in our models, as well as adding fixed effects for connective frequency and coherence relation at the trial level. We built iterative models beginning with a fixed effect of relation and compared successive models with additional predictors using likelihood ratio tests and AIC/BIC values. Factor levels were contrast (treatment) coded, with the baseline set to "low" for connective frequency, and "addition" for coherence relations. Our final model included significant main fixed effects of LexTALE ($OR = 1.24$, $p < .05$), AFT ($OR = 1.63$, $p < .001$), coherence relation, and frequency, and random intercepts for participants (Marginal-R^2: .19, Conditional-R^2: .24; Table 4.3.5). A model which included random effects for items did not converge, likely due to the overlap with the relation and frequency variables. Main effects for all coherence relations except "concession" and "consequence" were significant, although only "contrast" was positive ($OR = 2.06$, $p < .01$), meaning Korean participants were significantly more likely to accurately answer questions of this type compared to the "addition" relation. All interactions between relation and frequency were also positive, indicating higher frequency versions of all connectives were more likely to be correctly selected.

Table 4.3.5: Fixed effects of LexTALE, AFT, connective frequency, and relation, and an interaction between relation and frequency on odds of correct connectives selections; L1 Korean participants. Random intercepts for participants were also included.

	Correct		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	6.107	4.304 – 8.667	**<0.001*** **
LexTALE	1.236	1.049 – 1.455	**0.011***
AFT	1.363	1.148 – 1.618	**<0.001*** **
Relation [cause]	0.489	0.321 – 0.745	**0.001*** **
Relation [concession]	1.350	0.835 – 2.182	0.221
Relation [condition]	0.206	0.137 – 0.309	**<0.001*** **
Relation [consequence]	0.834	0.535 – 1.302	0.425
Relation [contrast]	2.062	1.219 – 3.489	**0.007****
Freq [high]	0.134	0.090 – 0.202	**<0.001*** **
Relation [cause] x Freq [high]	21.217	11.557 – 38.950	**<0.001*** **
Relation [concession] x Freq [high]	6.168	3.269 – 11.636	**<0.001*** **
Relation [condition] x Freq [high]	29.393	16.693 – 51.756	**<0.001*** **
Relation [consequence] x Freq [high]	15.507	8.165 – 29.451	**<0.001*** **
Relation [contrast] x Freq [high]	3.163	1.638 – 6.105	**0.001*** **
Random Effects			
σ^2	3.29		
$\tau_{00\ \text{Participant}}$	0.25		
ICC	0.19		
$N_{\ \text{Item}}$	12		

$N_{Participant}$	177
Observations	3480
Marginal R^2 / Conditional R^2	0.186 / 0.244

** p<0.05 ** p<0.01 *** p<0.001*

We also evaluated exploratory models with the interaction of L1 (i.e. language group), using treatment coding with English as the reference level, and French and Korean speakers as separate comparison levels, to determine if the print exposure measures (AFT/ART) were differentially associated with connectives scores as a function of L1. For this analysis, we opted to exclude the Oxford student sample, as they may represent a different native population when comparing performance on the outcome measures with the general UK population.

We iteratively added predictors to our model, comparing more complex models using likelihood ratio tests. Our final model (Marginal-R^2: .15, Conditional-R^2: .30) included fixed effects for L1, relation, frequency, and AFT, as well as interactions between L1 and frequency, relation, and AFT, and random intercepts for participants and items (Table 4.3.6). Here, we observed significant main effects for AFT ($OR = 1.30$, $p < .05$), showing print exposure was significantly associated with the odds of correct responses across language groups, and relation ($ORs =$ 2.64-5.84), indicating increased odds for all coherence relations compared to "addition". Although L1 was non-significant, its interaction with AFT showed the beneficial effect of print exposure was significantly higher for French speakers compared to English natives ($OR = 1.78$, $p < .001$), whereas the effect was equivalent between Korean and English natives ($p = .52$). Similarly, connective frequency alone was non-significant, but a positive interaction with L1 showed that

French natives were significantly more sensitive to frequency effects than native English speakers ($OR = 1.56$, $p < .01$), whereas the effect was equivalent between Korean and English natives ($p = .47$). Finally, the interaction of L1 on relation showed that both Korean and French participants were significantly less likely to correctly respond to each compared to English natives, although the relation of consequence was non-significant for L1 Korean speakers.

For comparison, an analogous model which replaced AFT with ART was similar overall, but with slightly lower Marginal-R^2 (.12 compared to .15), indicating moderately lower variance explained by the fixed effects. However, the main effect of ART in this model ($OR = 1.36$, $p < .05$) was comparable to AFT in the previous ($OR = 1.30$, $p < .05$), but did not show the same interaction with L1. This indicates that the effect of ART was similar across all language groups, which may be an advantage when comparing print exposure cross-linguistically. This contrast between the two models appears to reinforce the notion that the effect of AFT is particularly strong in French speakers, who may have more varied exposure to sources of English reading.

Table 4.3.6: Fixed effects of AFT, L1, connective frequency, and relation, and interactions for L1 on AFT, frequency, and relation on odds of correct connectives selections. Random effects of participant/item were also included.

		Correct	
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	2.567	1.226 – 5.376	**0.012***
L1 [FR]	0.831	0.547 – 1.264	0.387
L1 [KR]	0.820	0.539 – 1.248	0.355
AFT	1.299	1.041 – 1.620	**0.021***
Relation [cause]	5.839	2.247 – 15.174	<**0.001*****
Relation [concession]	3.738	1.458 – 9.583	**0.006****
Relation [condition]	5.806	2.234 – 15.086	<**0.001*****
Relation [consequence]	2.641	1.036 – 6.731	**0.042***
Relation [contrast]	3.399	1.328 – 8.697	**0.011***
Freq [high]	0.982	0.566 – 1.705	0.949
L1 [FR] x AFT	1.781	1.307 – 2.426	<**0.001*****
L1 [KR] x AFT	1.121	0.822 – 1.529	0.469
L1 [FR] x Relation [cause]	0.246	0.156 – 0.391	<**0.001*****
L1 [KR] x Relation [cause]	0.381	0.241 – 0.604	<**0.001*****
L1 [FR] x relation [concession]	0.608	0.392 – 0.942	**0.026***
L1 [KR] x relation [concession]	0.957	0.612 – 1.494	0.846
L1 [FR] x Relation [condition]	0.238	0.150 – 0.378	<**0.001*****
L1 [KR] x Relation [condition]	0.196	0.125 – 0.307	<**0.001*****
L1 [FR] x Relation [consequence]	0.466	0.308 – 0.707	<**0.001*****

L1 [KR] x Relation [consequence]	1.324	0.859 – 2.042	0.203
L1 [FR] x Relation [contrast]	0.323	0.212 – 0.493	**<0.001*****
L1 [KR] x Relation [contrast]	1.082	0.694 – 1.688	0.728
L1 [FR] x Freq [high]	1.562	1.197 – 2.040	**0.001****
L1 [KR] x Freq [high]	1.081	0.822 – 1.423	0.576

Random Effects	
σ^2	3.29
$\tau_{00\ Participant}$	0.56
$\tau_{00\ Item}$	0.20
ICC	0.19
$N_{\ Item}$	12
$N_{\ Participant}$	177
Observations	10620
Marginal R^2 / Conditional R^2	0.144 / 0.305

** p<0.05 ** p<0.01 *** p<0.001*

COLLOCATIONS

Our hypotheses stated that collocations scores would be positively predicted by LexTale (H1b), ART scores (H3a), and that the effect of ART would hold when controlling for LexTALE, but AFT would not (H3b). For collocations, the best-fitting linear regression model (Adj-R^2 = .44, p < .001; Table 4.3.7) included significant main effects for both LexTALE (β = 915.08, p < .001) and ART (β = 414.13, p < .05), confirming H1b and H3a respectively. Due to non-normality of residuals in this model, we applied a square power transformation on the

collocations outcome variable, which resolved the issue. Although AFT and ART were both equally correlated with the collocations variable ($r = .35$; Table 4.3.3), the effect of AFT was non-significant when controlling for either LexTALE or ART, confirming H3b. To illustrate the differential contributions of each main predictor, standardised partial residual plots from a model with all predictors are shown in Figure 4.15.

Table 4.3.7: Multiple regression model predicting collocations scores by print exposure measures (L1 Korean participants).

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2883.197	2563.324 – 3203.069	**<0.001*****
LexTALE	914.075	570.589 – 1257.560	**<0.001*****
ART	414.131	70.645 – 757.617	**0.019***
Observations	61		
R^2 / R^2 adjusted	0.457 / 0.438		
		** p<0.05* *** p<0.01* **** p<0.001*	

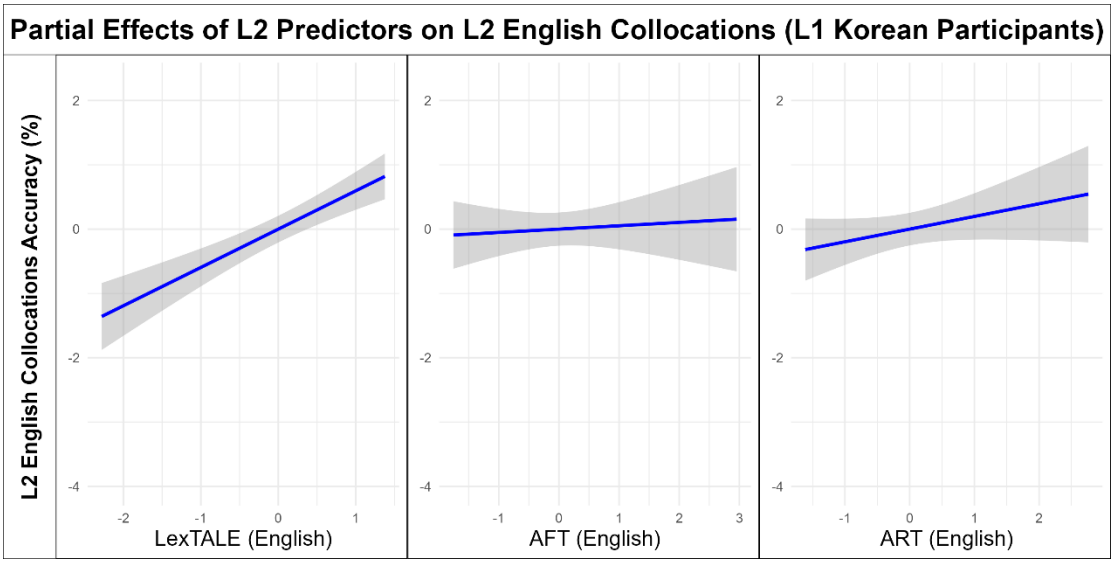


Figure 4.15: Standardised partial residual plots of all predictors on English collocations scores, L1 Korean participants. Shaded areas represent the 95% confidence interval.

We also evaluated exploratory mixed-effects models with main fixed effects for LexTALE, ART, and motivation, as well as their interactions with collocation frequency (using values from Dąbrowska, 2014), along with random intercepts for participants and items. Both ART and motivation were non-significant when accounting for LexTALE. Our best model showed a significant main effect for LexTALE ($OR = 1.67$, $p < .001$), and collocation frequency was marginal ($OR = 1.28$, $p = .06$). However, the low marginal-R^2 (.07) compared to the much higher conditional-R^2 (.27) suggested that most of the variance explained was from random factors.

4.4 General Discussion

Print exposure provides an important means of acquiring vocabulary which is less frequent in spoken language. However, the extent to which English learners rely on L2 reading as a source of second language knowledge is less clear, as are the means for assessing print exposure in these populations. In this study, we estimated print exposure (using AFT and ART) and language proficiency (LexTALE) for L1 Korean speakers of English using versions of each task in both languages, and evaluated how they were associated with L2 formulaic language knowledge (English connectives and collocations). We focused on three main research questions, namely i) whether L1 or L2 predictors are more highly associated with L2 vocabulary knowledge, ii) whether the L2 AFT or ART is more highly associated with L2 vocabulary knowledge, and iii) whether this depended on the vocabulary outcome measure.

We found that the L2 English LexTALE was positively correlated with L2 connectives and collocations, supporting H1a and H1b respectively. The L2 AFT was also significantly correlated with connectives scores (AFT $r = .50$), providing support for H2a. Additionally, the association between AFT and connectives remained significant when controlling for LexTALE in multiple regression models, providing partial support for H2b—although we had anticipated the ART would be rendered non-significant in a comparable model. Instead, we found that it was LexTALE which failed to explain additional variance in our models, and AFT and ART both independently contributed to knowledge of English connectives. Further, we found an interaction between the two print exposure measures, such that as scores on AFT increased, ART scores decreased, and vice-versa. Although this

interaction was not pre-registered, it reinforces findings from previous chapters suggesting the two measures may be complementary. That is, either the two measures are associated with different kinds of reading experiences, or they tap different memory resources—or both. In any case, it supports the assertion that the two measures assess print exposure through different means, and both may have their uses.

In contrast, both print exposure measures moderately but significantly correlated with collocations (both $r = 0.35$; see Table 4.3.3), supporting H3a, which anticipated that ART in particular would correlate with the dependent measure. However, linear regression models showed both print exposure measures were non-significant when controlling for LexTALE scores. This provided evidence for the null for H3b, which posited that ART would remain significant in this scenario.

Looking at L1 predictors, we found that none of the L1 Korean tasks predicted any L2 outcome variables, indicating a null role for L1 print exposure or proficiency. We conclude that we did not find evidence for either H4 (that LexTALE-KR would predict connectives/collocations scores) or H5 (that ART-KR scores would predict collocations scores, and would remain significant when controlling for LexTALE-KR).

To summarise in simpler terms, we found that for L1 Korean speakers, second language print exposure (as assessed by both AFT and ART) is significantly associated with knowledge of L2 formulaic language, but it seems to play a smaller role in their knowledge of English collocations compared to connectives.

What could explain this pattern of results? Firstly, we note that connectives, which encode a set of instructions for a clause's interpretation through coherence relations, are arguably easier to acquire through explicit learning compared to collocations—the latter being nearly infinite, and arbitrary almost by definition. Intuitively, we might suspect that collocations require more "ambient" exposure to print in the environment, which may not be as accessible for Korean natives living overseas (compared to, e.g., French speakers). It may also be due in part to collocations being more congruent (i.e. similar or equivalent words tend to appear together cross-linguistically) between French and English compared to Korean and English. Appealing though they may be, these intuitions appear unlikely, as we found that Korean and French native speakers' performance was not statistically different on the English collocations task (Table 4.3.1). Evidently, Korean speakers are keeping pace with the geographically and typologically closer French speakers, although both predictably trail behind English natives.

As a language proficiency measure, it is unsurprising that LexTALE is highly correlated with L2 vocabulary. Yet in the previous chapter (McCarron et al., 2025), we still saw a marginal effect of AFT on collocations scores when controlling for LexTALE in the French population, showing print exposure is likely a unique contributor to formulaic vocabulary learning in L2. We expected we might see the same result here as well, but this was not the case. Granted, LexTALE is its own kind of proxy for other proficiency measures, and proficiency and print exposure are generally understood to correlate reciprocally due to Matthew effects for reading, making their unique contributions difficult to disentangle. In retrospect then, the pre-registration's specification that the effect of print exposure should

remain significant even when accounting for general L2 proficiency may have been overly stringent, or even unnecessary to interpret the role of print exposure for vocabulary learning.

With respect to language transfer theories, we anticipated that opportunities for learning English may be more constrained by existing native language proficiency in South Korea. However, we saw no evidence for an effect of L1 Korean proficiency (LexTALE-KR) or print exposure (AFT-KR, ART-KR) on L2 formulaic language outcomes. Granted, the relationship may be more indirect than we have accounted for presently—for example, L1 proficiency may serve as a mediator to L2 proficiency, which in turn affects L2 vocabulary outcomes. This may seem unlikely, given that there was no correlation between the L1 and L2 LexTALEs. However, a structural equation modelling approach would have allowed us to test for these indirect relationships while also including potential underlying latent variables such as cognitive ability or language aptitude, which may influence proficiency in both languages. Without these analyses, the current results may have little direct bearing on the research literature surrounding language transfer.

Unlike the previous study, we did not measure general semantic fluency using categories such as animals or public figures. Logically, this skill may be a mediating factor in AFT performance, as an individual's score is dependent on speed of lexical access. However, the previous study showed that semantic fluency mediated the effect of AFT scores on connectives, but not collocations. This also varied by subtask, with the proper nouns (public figures) having no relationship with either outcome variable. This is important, because authors and public figures are both kinds of proper nouns, meaning they share similar lexical access pathways,

and the lack of an effect for public figure fluency indicates that print exposure is the limiting factor (see previous chapter for more discussion on this topic). Consequently, we suspect the inclusion of a general semantic fluency task would have provided little additional insight.

One of the fundamental (though as yet unproven) assertions in favour of using a semantic fluency measure like AFT is that the names provided are potentially more likely to represent explicit memory of reading experience compared to ART. In other words, these names should be more indicative of primary as opposed to secondary print exposure. As mentioned in the introduction to this chapter, some sources have claimed that Korean speakers are perhaps less likely to have encountered author names through primary print exposure (H. Lee et al., 2019). In support of this, H. Lee and colleagues asked which ART author selections participants had read personally, and on average this figure was considerably lower than the total number of names selected. Ideally, this study would have done the same for names provided on both print exposure measures. As mentioned previously however, the process of cleaning and separating names on AFT is often laborious and time-consuming, and it may not have been feasible in an online setting to analyse, clean, and store these names for retrieval without any kind of manual intervention.

In sum, compared to the results of the previous chapter suggesting AFT is more useful than ART in L1 French speakers of English, the findings here are more equivocal. Fundamentally, either print exposure measure performs similarly well as a predictor of L2 formulaic language in a Korean-speaking population, with no strong evidence to support the use of one over the other. Regardless, given the

ubiquity of ART as a proxy measure of print exposure, it is surprising that an open-ended semantic fluency measure of print exposure performs as well or better in L2, depending on the population. With this in mind, we recommend the use of AFT in place of, or as a supplement to ART as a measure of print exposure in L2 populations.

4.5 Data Availability

Data and code used in analyses are available on OSF (https://osf.io/3p6vh/).

5. Effects of L2 reading experience on predictive processing of spoken idioms

> "What, then, we may ask next, is the proper use of words? Not, so we have said, to make a useful statement; for a useful statement is a statement that can mean only one thing. And it is the nature of words to mean many things. [...] Thus one sentence of the simplest kind rouses the imagination, the memory, the eye and the ear — all combine in reading it."
>
> — Virginia Woolf, 29 April 1937 (1942, 2012)

5.1 Introduction

5.1.1 Predictive Processing in L1

Consider the problem of parsing language into meaningful units. From infancy, children must acquire a phonemic inventory of sounds in their target language, learn to segment speech to demarcate word boundaries, and determine what these words refer to in the world (Bloom, 2000, 2001, 2005). This learning of the referential meaning of words is referred to as the "mapping problem" (Carey, 2010;

Trueswell et al., 2013). In broadcast journalism and film editing, a common rule of thumb is often expressed in the maxim, *"say dog, see dog"*—in other words, audio and visual information should match, even if this is rarely the case in practice (Friez, 2022). Similarly, parents and caregivers may try to pair the name for an object with repeated presentations of the object itself, but these associations are unreliable and are often insufficient for determining a word's referent (Bloom, 2000). Apart from these special circumstances in which a sign and its signifier are explicitly joined, children do not acquire language word-by-word, building a lexical repertoire like stacking their LEGO bricks. Instead, an accumulation of evidence shows that children begin by processing language at the phrase level (e.g., Conklin & Schmitt, 2012; Foster, 2001). Under these "constructivist" or "usage-based" accounts of language development (Bannard & Matthews, 2008; Bybee, 2010; Dąbrowska & Divjak, 2019; N. C. Ellis, 2002; Tomasello, 2010), children first understand the communicative intent of a message and its context, and clearer representations of individual words take hold with time. To continue the previous analogy, we might say that this is comparable to children picking up a completed LEGO set of a car and playing with it, before later realising that the wheels can be removed and repurposed for any number of other conventional configurations—not just cars, but skateboards, aeroplanes, or lunar rovers. As we will see, this may have important implications for understanding the differences between how native and non-native speakers typically learn the same language (Arnon & Christiansen, 2017; Conklin & Schmitt, 2012; N. C. Ellis, 2002).

Whether native or non-native, however, speakers of any language have a remarkably narrow window of time to parse and process an incoming message, be

it written or spoken. The ever-present limitations of memory mean that each sentence can only be stored temporarily, especially as new information streams in continuously, inexorably. This is particularly evident in speech, as the audible facsimile in the working memory's phonological loop quickly fades (Baddeley, 1972; Baddeley & Hitch, 1974). In comparison, writing has the considerable advantage of external storage, providing a durable record which readers can refer back to. (On the other hand, crack open a stream-of-consciousness, avant-garde novel with winding run-on sentences, and note just how frustrating it is to try to integrate the end of a sentence with the beginning.) Christiansen and Chater (2016) refer to this narrow window for language processing as the "Now-or-Never bottleneck". Essentially, in order to accurately and efficiently process language input before it vanishes from memory, the brain "recodes and compresses" information into "chunks", extracting a multilayered representation of its semantic structure and intended message, which is then passed along to higher levels of processing. From there, the brain quickly moves on to the following sentence to do the same ("chunk-and-pass" processing)—lather, rinse, repeat. Unsurprisingly then, if you ask someone to recall verbatim the contents of a conversation from earlier the same day, you're bound to be disappointed—instead, you'll hear a broad stroke account of the general tenor or sentiment of the exchange. Yet this is the language faculty working as intended, as this shallow level of processing also enables it to work efficiently (Christiansen & Chater, 2022).

Of course, speakers do not simply listen passively as a speech stream washes over them—or as their eyes wash over the words on a page. Rather than waiting for the end of a sentence or a conversation turn to begin interpreting a message, a

wealth of eye-tracking and electroencephalographic evidence indicates that language users play an active role, building a structural representation incrementally and anticipating what is likely to follow (dubbed "predictive processing"; see Kaan, 2014; Kutas et al., 2011, and references therein). This anticipatory behaviour is not unique to language, and may in fact be fundamental to human cognition generally (A. Clark, 2016; Williams, 2018). Under a predictive processing framework, the mind constantly generates expectations about what will happen next, using past experience and present context to build a probabilistic model of the world. This process is also dynamic, with evidence from eye-tracking and self-paced reading studies showing that sentence and word reading interact to jointly influence prediction (Amenta et al., 2023).

If predictive processing is useful for language comprehension, it may be worthwhile to understand what makes a "skilled predictor". Some have argued that prediction is facilitated by language production, through a form of internal imitation (e.g., Pickering & Garrod, 2007, 2013). For example, evidence suggests that semantic prediction (e.g., *eat* predicts *cake;* Altmann & Kamide, 1999) is related to productive vocabulary rather than comprehension skill (e.g., Mani & Huettig, 2012). Evidence of this kind has led Pickering and Garrod (2013) to propose an "integrated theory of language production and comprehension". Under this theory, Pickering and Garrod posit that prediction is essential and fundamental to language processing, and that each speaker builds a "forward model" of each sentence, anticipating what comes next with every new piece of information. To return to the Altmann and Kamide example, if a child were to hear the word "eat", based on her internal production of the sentence, she may be more likely to predict

an edible noun such as "cake" compared to something non-edible (such as "shoe"), as the transitive verb "eat" triggers a selection criterion requiring it to be followed by a noun phrase (a syntactic component) which must also be an edible object (a semantic constraint).

Chang's (2002) connectionist "dual-path" model of sentence production (which proposes that meaning and sequence/structure are processed separately) similarly suggests that production and comprehension are interrelated aspects of language competence. This related theory also supposes that speakers develop an internal model of sentences during online processing, but this model is updated when these predictions prove to be incorrect (i.e. "error-based" learning). Unlike Pickering and Garrod's theory then, the dual-path model explains predictive behaviour as a consequence, rather than a cause, of learning (Chang et al., 2013). Because of this, prediction is thought to be employed more often in the acquisition of language as opposed to its processing.

In any case, it is clear that production, comprehension, and prediction are all related to the efficient processing of language. Naturally, all of these factors are partly contingent upon experience, and as detailed in previous chapters, learning to read is a particularly important source of language exposure (Bloom, 2005). Yet, evidence shows that L1 literacy trains readers not only to more effectively process written language, but to predict speech as well (Favier et al., 2021; Huettig & Pickering, 2019). This demonstrates that the experience acquired through reading does not remain restricted to the written mode, but rather this pool of language experience can be drawn upon in related domains.

5.1.2 Predictive processing in L2

Compared to natives, L2 speakers do not anticipate upcoming language input quite so readily. Formalising this observation, the RAGE hypothesis states that L2 speakers demonstrate a "Reduced Ability to Generate Expectations" during online language processing (Grüter et al., 2014). Much of the evidence for this comes from eye-tracking technology and visual world paradigms, a combination which has proven to be an important tool for psycholinguistic investigation (Huettig et al., 2011). Visual world paradigms are an experimental task design where images are presented in different areas of the screen (typically four quadrants) and eye movements are recorded during online processing of a sentence to determine where participants are looking and when. Although much early work of this kind looked at native speakers (Altmann & Kamide, 1999; Cooper, 1974), a significant body of research has demonstrated that it can be an effective means of inquiry in adult L2 populations as well (Dussias, 2010). For example, compared to natives, L2 speakers are slower to predict semantic information (Dijkgraaf et al., 2019), phonological forms (Ito et al., 2018), and pronoun referents in verbs with implicit causality (Kim & Grüter, 2021). Yet it has been argued that these findings are not due to some essential nature of L1 and L2 speakers (as might be predicted under the "fundamental difference" hypothesis; Bley-Vroman, 1989, 1996), but rather that identical causes contribute in both groups (e.g., individual differences in lexical quality, item frequency, etc.; Kaan, 2014). Although L2 speakers do understand grammatical rules about their target language when asked about them explicitly, they show greater difficulty with online prediction than L1 peers (Kaan, 2014, and references therein, see esp. p. 259). This aligns with work by showing L2 speakers

perform similarly to natives on comprehension questions, but struggle most on fluency measures (Kuperman et al., 2023), suggesting L1/L2 differences are primarily a question of degree rather than kind (Siegelman et al., 2024). In other words, rather than swimming upstream against the current of linguistic parameters which were set in L1 under a universal grammar framework, L2 speakers may simply not have sufficient relevant experience (Arnon & Christiansen, 2017; Conklin & Schmitt, 2008; Kessler & Beck, 2022). In the following sections, we elaborate on the motivation for our current study which employs a visual world paradigm to evaluate predictive processing of idioms in an L2 population.

5.1.3 Formulaic and idiomatic language

Because of the interest in usage-based theories of language acquisition which posit that meaning may begin at the phrase level, a substantial body of work has examined the processing of formulaic language (sometimes called multi-word expressions or multi-word units) in recent decades (Wray, 2002, 2006). Formulaic expressions can be useful for chunking language into larger units, which eases the processing burden between communicators (Siyanova-Chanturia, Conklin, & Van Heuven, 2011); correspondingly, such phrases are processed more quickly, both in L1 and L2 (Conklin & Schmitt, 2008). Examples of multi-word expressions include fixed binomials ("bride and groom" vs "groom and bride"), collocations ("raise prices", "torrential downpour"), discourse fillers (especially in speech, e.g. "you know", "a lot of", "at the end of the day"), phrasal verbs ("give up", "get up", "turn

up"), proverbs (conventional wisdom such as, "the early bird catches the worm"), and idioms.

Idioms are special kinds of formulaic phrases for which the meaning cannot be wholly determined from its constituent elements—for example, to "kick the bucket" (to die) or to be "skating on thin ice" (to be in a precarious or dangerous situation). Idioms vary in terms of their semantic decomposability (or "transparency"); that is, the degree to which meaning can be recovered from the individual words themselves (Libben & Titone, 2008). For example, a novice French learner can be reasonably certain that a novel expression such as *« passer une nuit blanche »* (literally, "to spend a white night") must be related to how someone slept the previous night (not well, unfortunately), whereas *« être roulé(e) dans la farine »* ("to be rolled in the flour") may not immediately reveal the phrase's figurative meaning (that someone has been tricked or scammed). In this way, learners must develop sensitivity to when words are being used literally, and when meaning can only be constructed through a figurative interpretation, which may require some degree of metaphoric inference and the chunking of multiple words into a discrete unit. Idioms have long been considered to be stored and retrieved from the lexicon as units, much like single words (Swinney & Cutler, 1979). It is tempting to assume that at a certain point in processing an idiomatic phrase, a threshold of evidence is met (i.e. the required constituent elements of an idiom are present), at which point a figurative interpretation takes hold and the literal meaning is discarded. Yet Beck and Weber (2016) reported that literal and figurative meanings of idioms are activated concurrently in both L1 and L2 speakers, meaning a listener will likely be primed for both "bucket" and "death"

when hearing a phrase such as "kick the bucket". The question is how quickly the phrase is chunked and packaged into a meaningful idiomatic interpretation, and the literal interpretation is suppressed, if not fully discarded. Given that L2 speakers tend to process language serially (i.e. "word-by-word") compared to natives, who generally acquire language in larger chunks (Conklin & Schmitt, 2012; Foster, 2001), there is some evidence to suggest that non-natives prefer literal interpretations of idiomatic phrases in L2 (Cieślicka, 2006; Siyanova-Chanturia, Conklin, & Schmitt, 2011), although this is contradicted elsewhere (Beck & Weber, 2016).[19]

In the previous chapters, we investigated the utility of the Author Fluency Task (AFT) as a proxy measure of print exposure, using measures of formulaic vocabulary (L2 connectives and collocations) as outcome variables. Although the connectives task required participants to incorporate words into a sentence frame, fundamentally each question measured vocabulary skill at the single word level; for example, understanding the contrastive function of a connective such as "whereas". Yet there is increasing acceptance that restricting the measurement of vocabulary knowledge to the word unit is arbitrary, as meaning is parsed at multiple levels, including morphology, words, and phrases—moreover, they may be processed by the same underlying cognitive system ("single-system" models of language; Arnon & Christiansen, 2017; Christiansen & Chater, 2016). Since most L2 teaching

[19] Anecdotally, as a fledgling French speaker I once heard some friends discussing a televised French-Canadian game show for secondary school students called *« Génies en herbe »* (*« herbe »* meaning grass, but *« en herbe »* meaning "budding", an expression which I did not know—in this case, the full title translates best to "Budding Geniuses"). Confused, I embarrassed myself by asking *« Est-ce une émission sur le jardinage ? »* ("Is that a show about gardening?")

generally focuses on single words, multi-word units may provide a better index of an L2 speaker's experience with words in context.

In contrast to the connectives task then, the *"Words That Go Together"* task (Dąbrowska, 2014) measured multi-word (bigram) knowledge. These word pairings can appear to be arbitrary, in the sense that many different combinations might convey the same literal message. Yet some pairings are more common than others, and thus sound more appropriate as well. Naturally then, collocations require a great deal of statistical learning from input, but speakers must be able to determine which elements are meaningful rather than simply those which tend to co-occur. The distinction between these two notions is illustrated using a mutual information (MI) index, developed in the fields of probability and information theory. Two variables (say, for example, two separate words) are said to have a high MI score if they co-occur more often than would be expected by chance when considering their individual frequencies (Church & Hanks, 2008; Dąbrowska, 2014; Durrant & Doherty, 2010). Unlike raw frequency, which only indicates how often a given collocation appears in any given corpus, MI indicates word pairings which may be less frequent than others, but which are more meaningful. To illustrate, a phrase such as "good morning" is high frequency because both words individually are high frequency. Comparatively, a bigram pairing such as "raise prices" may not appear nearly as often—when it does however, its union is anything but arbitrary, as this particular combination of words is preferred compared to alternatives (e.g. "lift prices"). Accordingly, research shows that whereas L2 processing speed is determined by a collocation's frequency, L1 processing is influenced more by a collocation's mutual information index (N. C. Ellis et al., 2008). Non-native

speakers who are still developing familiarity with the target language, evidently focus more on tracking simple frequencies, rather than the deeper knowledge of how surprising a particular combination of words is.

Additionally, the measures used in the previous chapters were intended to measure explicit, offline judgements of language knowledge. Yet L1 and L2 speakers are primarily distinguished in terms of fluency rather than comprehension (Kuperman et al., 2023), and show signs of implicit knowledge of their L2 which is not always readily accessible through explicit recall (R. Ellis et al., 2009; Zufferey et al., 2015). Consequently, L1/L2 differences might be most evident through implicit online measures such as eye tracking, which can gauge anticipatory processing of language. Finally, despite ample research indicating prediction is an important aspect of language processing and comprehension, and recognition that individual differences and experience are involved in predictive behaviour, there is apparently no work linking print exposure to predictive processing in L2.

5.1.4 Present study

To briefly summarise the preceding evidence, children seem to learn their first language by learning phrases, whereas second language speakers typically start by learning individual words, often through direct translation from L1. For this reason, formulaic language may be a particularly fruitful avenue for assessing exposure to print in L2, as knowledge of these longer forms may suggest greater naturalistic exposure to the target language. Additionally, some theories position predictive processing as essential to all aspects of human cognition, including language.

Essentially, these theories state that the brain builds a model of the world around it, and this model updates according to the strength of evidence. Although language prediction and comprehension are correlated, the direction of causality is debated. Finally, some theories have suggested that predictive processing of language relies upon a system of internal production—that is, a reader or listener generates a model of the sentence in real time to determine what is most likely to follow. Consequently, for the purposes of the assessment of print exposure, we reason that it is possible that a generative production task such as AFT will be more highly associated with predictive processing of language than a recognition task such as ART.

In this study, we sought to determine if print exposure is associated with predictive processing of formulaic language in L2. To do so, we needed to design a novel task evaluating L2 speakers' online processing of sentences using formulaic phrases where multiple words must be integrated into a meaningful unit—in other words, "chunking" bits of language together. Because many idioms are non-decomposable, we reasoned that L2 speakers would require greater print exposure to be able to quickly understand which sentences are idiomatic as opposed to literal. Using a visual world paradigm and eye fixation counts, we could determine how often L1 and L2 speakers look to iconic images which represent the semantic contexts for either a literal or idiomatic interpretation. Because we looked at L1 French speakers of English in a previous chapter, we decided to sample the same population for this study. However, recruiting a large sample of native French speakers on-site was not considered feasible due to the relatively few available speakers in proximity to Oxford. In recent years, however, advances have made it

possible to run eye-tracking experiments over the Internet using participants' home web cameras (Prystauka et al., 2023). For these reasons, we opted to run this study online using custom scripting based on WebGazer.js (Papoutsaki, 2015; Papoutsaki et al., 2016) through Gorilla Experimental Builder (Anwyl-Irvine et al., 2020).

The aims for this study were both methodological and theoretical. Methodologically, we aimed to determine the suitability of Internet-based eye-tracking measures. Although this is not the first study of this kind, to date there are relatively few (e.g., Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2023), as the nascent technology has only existed for a few years. Additionally, we wanted to develop and validate a novel measure for assessing knowledge of idioms using a visual world paradigm. To make this possible, we would also determine how image generation software using large language models (i.e., so-called "artificial intelligence") can be used to facilitate the production of iconic images for psycholinguistic studies.

From a theoretical perspective, we asked if reading experience would transfer between written and spoken modes in L2 as well as L1 speakers, and how L2 speakers can learn to "chunk" linguistic information as a function of print exposure. Using the iconic images which represent different semantic contexts, we can also determine if reading experience facilitates the processing of visual information in addition to spoken language. Thus our research questions focused on how print exposure generally correlates with predictive processing of L2 formulaic language.

Our hypotheses were as follows:

H1: If L2 reading experience affects predictive processing of speech, then participants with greater L2 print exposure (as indexed by AFT) would make increased anticipatory looks to target images for idiomatic sentences.

H2: If L2 reading experience affects knowledge of idiomatic language, then participants with greater L2 print exposure (as indexed by AFT) would have higher VWP image selection accuracy for idiomatic trials.

In contrast to the previous chapters, we did not initially intend to compare the two measures of print exposure (AFT and ART). Because we included both tasks in our experimental battery however, we contrast the two methods in terms of their predictions on an exploratory basis.

This study received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee [reference R77364/RE004] and was pre-registered on OSF (https://osf.io/ec98w).

5.2 Methods

5.2.1 Participants

We recruited 50 L1 French speakers who spoke intermediate to advanced English as L2 (20 women) who were aged between 22 and 68 years of age (M = 40.13, SD = 14.82). For our comparison sample, we recruited 50 L1 UK English speaking

participants (25 women) who were raised monolingual and were aged between 21 and 63 ($M = 41.84$, $SD = 10.43$).

Prior to data collection, power analysis was conducted using G*Power version 3.1.9.7 (Faul et al., 2007). Using a standard .05 alpha error probability with four predictors, this software recommended a sample size of 45 for 0.8 power to detect a medium effect size of .3. In order to recruit a reasonably large sample of L1 French speakers, the decision was made to carry out the study online with eye tracking software which uses participants' computer webcams. Participants were recruited through Prolific (*Prolific*, n.d.) and tested in one session using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020), although a smaller subset returned to complete some of the measures again in a second session, in order to calculate test-retest reliability.

5.2.2 Measures

At the conceptual stage, we began by listing "imageable" idioms—in other words, expressions for which both the literal and the figurative meanings were sufficiently concrete to render as representational drawings. It was equally important that each idiom should not be "decomposable", that is, it should not be possible to derive its meaning from any constituent parts (Libben & Titone, 2008). We also consulted a database of English idiom norms (Bulkes & Tanner, 2017), seeking to include idioms with moderate to high ratings for familiarity and literal plausibility, although not all of the expressions we selected had norms available. From an initial list of 150 English idiom candidates, we selected a total of 60. Of these, 44 were in

the Bulkes and Tanner database of norms, with an average familiarity rating of 3.40/5 and an average literal plausibility rating of 3.20/5.

Using corpus query language (CQL) in *SketchEngine* (Kilgarriff et al., 2014), each idiom was queried in both the written and spoken subcorpora of the British National Corpus (BNC) to determine how frequently each idiom is found in writing relative to speech. However, due to a significant number of instances where an expression was not found in the spoken corpus, we instead decided to record the frequency in the complete BNC to provide a general measure of frequency, which we use in later analyses.[20] This left us with 57/60 phrases with BNC frequency values, which had an average of 0.13 occurrences per million ($SD = 0.12$). For a full list of sentences and corresponding CQL search strings, please see Appendix A.14.

After deciding which idioms to use, each phrase was embedded in a sentence which began with a subject noun followed by the idiom, followed by 5-12 words which brought the sentence to its conclusion. These additional words were primarily intended to extend each sentence's length, providing participants more time to make predictive looks, so we took care to avoid providing additional details about the meaning of the idiomatic phrase until the final critical word or phrase. An example idiomatic sentence is provided in example 1:

[20] However, we note that the average ratio of written to spoken frequency values (available for 26/60 phrases) was 0.997 ($SD = 0.850$), whereas the average ratio of written to full BNC frequency values (available for 56/60 phrases) was 0.992 ($SD = 0.170$). We suspect that this broad equivalence means that the ratio may not have been as informative a measure as we originally hoped, and the frequency value is likely more useful.

1) He finally *kicked the bucket*,
but I couldn't believe he was really *dead*.

Analogous literal versions of idiomatic sentences were written to provide comparisons between conditions, see example 2:

2) He finally *filled the bucket*, but
before doing so, he lost a lot of *milk*.

Literal trials featured the same images for the visual world paradigm as their idiomatic counterparts, but the target and distractor images were switched. Participants would see one of either A or B versions of the stimuli, such that each participant would see an equal number of literal and idiomatic expressions, but never the same images twice on separate trials. As an example, a hypothetical Participant 1 would be assigned to Version A, and they would hear the idiomatic *"kicked the bucket"* sentence, and their target would be the image of the grave. In contrast, Participant 2 would be assigned to Version B, and would hear the literal *"filled the bucket"* sentence, and their target would be the image of the bucket. See Figure 5.1 for an illustration of how targets and foils varied from idiomatic to literal trials.

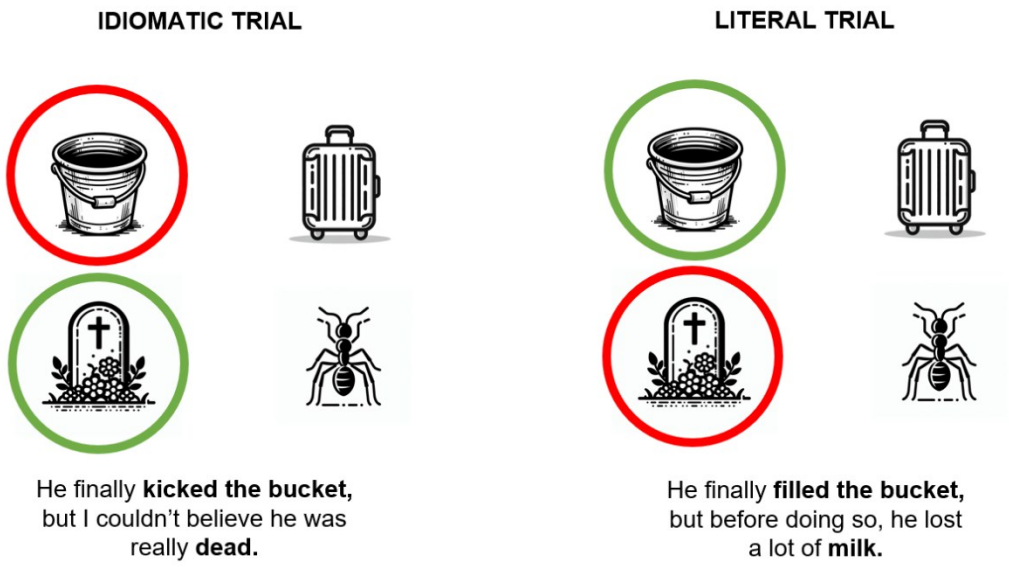


Figure 5.1: Illustration comparing idiomatic and literal trials. Green circles show the target image, and red circles show the foil image. Each participant was presented only with the idiomatic or the literal version of each sentence, but saw an equal number of both conditions. On the actual trials, image locations were also randomised, but are presented here in identical locations for interpretability.

Each sentence began with a noun or pronoun subject followed by an idiomatic or literal phrase. Because each sentence varied in length, we coded the onset timings in milliseconds for cues, critical words, and sentence completions for each sentence recording. These timings enabled us to restrict our analyses for predictive looks-to-targets for each sentence individually. Across all trials, average cue onset was 717 ms (figurative = 735, literal = 704), average critical word onset was 3697 ms (figurative = 3752, literal = 3642), average sentence duration was 4463 ms (figurative = 4483, literal = 4444), and average predictive window length was 2978 ms (figurative = 3017, literal = 2940).

Designing a visual world paradigm of this size (126 trial and practice sentences in total) required the creation of 264 novel images. To facilitate this task, we used image generation software DALL·E 3 (OpenAI, 2024) which is built on the technology of large language models like ChatGPT. Natural language prompts were given using the following frame (3):

3) *"Simple black and white line-art cartoon drawing of a _____. Solid white background, no shading."*

For examples of these images with accompanying idioms, see Figure 5.2. Of these generated images, the overwhelming majority required some degree of manual intervention to correct minor rendering errors (e.g., a light grey or off-white background rather than solid white), or to better convey the intended meaning of each sentence (e.g., adjusting character expressions to more clearly signal a particular emotion). In some instances, content safety filters prevented the rendering of images as described—for example, an image of a knife representing the literal phrase *"he arranged the killing",* for which exaggerated blood needed to be drawn separately. In other cases, I had to composite multiple generated images to create the intended scenario, as in the images representing "shame" (a man looking ashamed as a woman looks at him with disdain) and "illness" (a man with a runny

nose with a box of tissues behind him). I drew 7 of the images which could not be rendered adequately by the image generation software.[21]

Figure 5.2: Example iconic images produced for the visual world paradigm using DALL-E 3 image generation software (OpenAI, 2024). For the full index of image stimuli, see Appendix A.15.

To determine the suitability of target images for each sentence, pilot data was collected from 4 L1 and 3 L2 volunteer raters. The average accuracy scores (calculated as the percentage of correct image selections) were 99.01% in L1 and 96.45% in L2, suggesting the intended contexts for the images were clear.

[21] For reference, the images which I drew were: "calendar.jpg", "ear.jpg", "fingers_crossed.jpg", "intersection.jpg", "knot2.jpg", "moon.jpg", "pool.jpg". See Appendix A.15 for the full list of images and corresponding file names.

Accordingly, we evaluated the few cases of mistaken selections and determined there was no consistent pattern which might indicate that a particular image caused confusion. Sentence recordings were generated using Microsoft's Azure Text-to-Speech software (Microsoft, 2024), using the "Sonia (English-UK)" voice, and all recordings were validated for authenticity by a native speaker of UK English.

For detailed descriptions of the print exposure measures used in this study, the Author Fluency Task (AFT) and Author Recognition Test (ART), see Chapter 2, and see Chapter 3 for descriptions of the English LexTALE, the Demographics Questionnaire, and Semantic Fluency Task. In this study, we collected semantic fluency performance for grocery store items and public figures in order to have examples of both common and proper noun fluency. We opted to leave out the animal fluency task used in the previous study as a time-saving measure, and because it may have been redundant, given that both animals and grocery items are common nouns. In all other instances, each task's instructions, scoring logic, and procedures were identical to prior iterations.

5.2.3 Procedure

Participants began by providing informed consent. Next, they were introduced to the eye tracking task and were presented with instructions for ensuring proper tracking using their webcams. Participants' webcam polling rates were then evaluated to ensure they relayed a minimum of 26 frames per second. If successful, participants proceeded to the calibration and validation screens, in which they were instructed to look at different points on the screen while eye position was recorded.

Upon successful validation, participants continued to a check of their audio equipment, where they could repeatedly play a sample audio file using the same voice as the subsequent experiment to ensure they could hear the sentences comfortably. Participants were then automatically assigned to either the A or B version of the eye tracking task (which determined the version of each sentence trial they would hear), ensuring an equal number of participants were assigned to both. Prior to beginning, participants were instructed that during each trial, they would see four images on screen and hear a sentence. They were advised to click on the image which best corresponded to the context described in each sentence. Prior to the main trials, six practice trials (three literal, three idiomatic) were shown for each participant. Correct/incorrect feedback was provided for each practice selection, as well as a brief explanation for the meaning of each sentence, but no feedback was provided for the main trials.

The order of individual trials was randomised, and the locations of images on screen were randomised per participant and trial. First, a fixation cross appeared in the centre of the screen for 1000 ms. Next, four images appeared during a "preview window" of 2000 ms and would remain on screen for the remainder of each trial. These preview windows were intended to allow participants to familiarise themselves with the semantic associations for each image.[22] Next, each sentence was played aloud. Finally, participants were prompted to select an image. Selection

[22] Our rationale for this preview window length was that it would allow time for participants to familiarise themselves with the images and their associated semantic meanings, but would likely be insufficient to develop strategies or make predictions (Berends et al., 2016).

accuracy and eye movement data were recorded for each trial. Trial structure is illustrated in Figure 5.3.

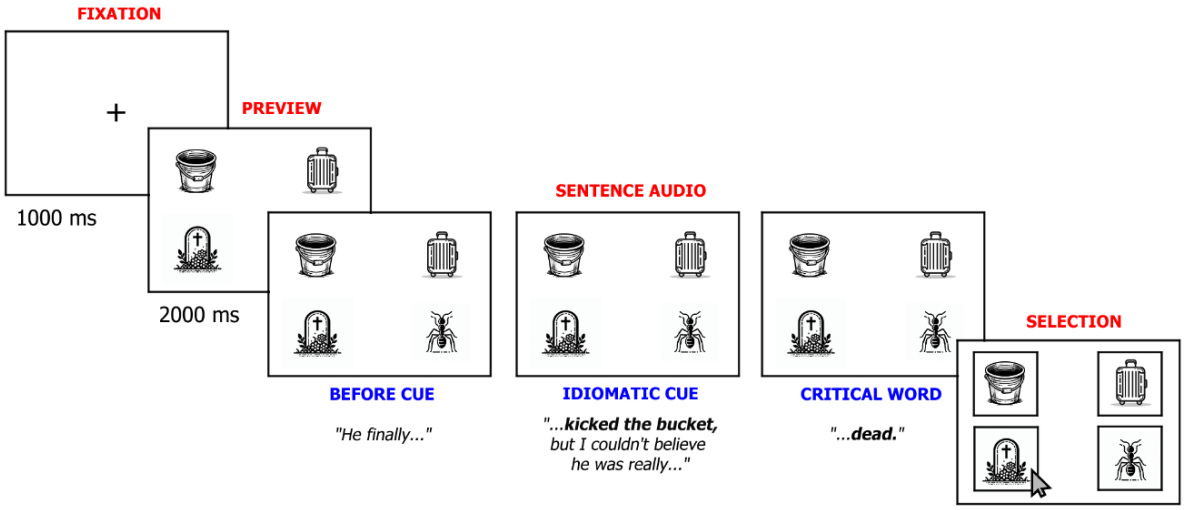


Figure 5.3: Example illustration of the visual world paradigm trial structure. Labels in red indicate the separate screens for each trial of the visual world paradigm, and labels in blue indicate the different sections of the sentence audio for processing. Analysis focused on the region in which the idiomatic cue was read, prior to the onset of the critical word.

Following the eye-tracking task, participants completed the English print exposure measures (AFT, ART), the English LexTALE (Lemhöfer & Broersma, 2012), and the demographics survey.

5.3 Results

We began by observing distributions of our measures and calculating summary statistics for each. Summary statistics for each predictor are shown in Table 5.3.1, and for outcome variables in Table 5.3.2. Table 5.3.3 shows a correlation matrix for all tasks in the L1 sample, and Table 5.3.4 shows the same for the L2 sample.

5.3.1 Predictor Variables

In this section, we describe performance on each of the predictor tasks used by language group (L1/L2), and performance distributions for each task by group are illustrated using raincloud plots in Figure 5.4.

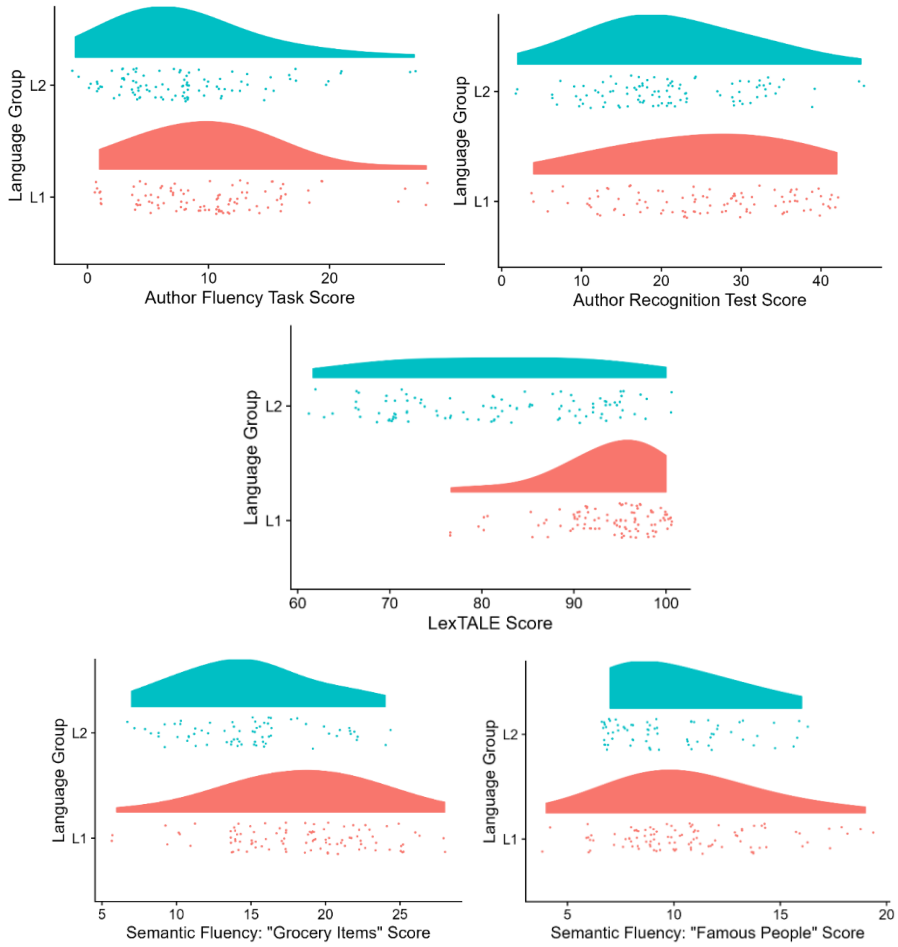


Figure 5.4: Raincloud plot illustrating distributions and observations of predictor task scores by language group.

AFT scores ranged between 1 and in 28 L1 ($M = 10.16$, $SD = 5.75$), and -1 and 27 in L2 ($M = 7.96$, $SD = 5.73$), see Table 5.3.1.[23] Score distributions by language group are illustrated in the raincloud plot in Figure 5.4. We calculated test-retest

[23] Although rare, negative AFT scores were possible if a participant were to provide the name of someone who is not a writer (e.g. one listed the actor Adam Sandler) or not primarily known for their writing (e.g. Barack Obama). These names were scored -1, whereas names which could not be verified were given a score of 0.

reliability of the AFT by inviting participants to return to complete the task again. In the L1 sample, 45/50 participants returned, obtaining a good reliability between time 1 and time 2 ($r = 0.80$, SEm = 2.55). TRR for individual author names and their counts from time 1 and time 2 (where a given name appeared at least once each time) showed a moderate correlation (Spearman's $r = 0.59$).

In our L2 sample, 34/50 participants returned, obtaining a moderate test-retest reliability (Spearman's $r = 0.68$, SEm = 3.22). Although this correlation is modest, this may be due to the lower sample size compared to the L1 group. However, this test-retest reliability value may not be abnormal, as it is identical to one found by Harrison et al. (2000) in a sample of 90 British participants on a similar semantic fluency task, and also corresponds with Lin and Chih (2023), who report moderate to high reliability ($r = 0.50$–0.80) in their study using various 60 second semantic fluency tasks in a non-clinical sample of L1 Chinese speaking Taiwanese adults ($n = 58$). Finally, reliability for individual author names and their counts from time 1 and time 2 showed a similarly moderate correlation (Spearman's $r = 0.64$), suggesting that although the pool of names drawn from memory are relatively stable, they do fluctuate somewhat.

ART scores ranged between 4 and 42 in L1 ($M = 25.62$, $SD = 10.27$), and 2 and 45 in L2 ($M = 20.92$, $SD = 9.04$), see Table 5.3.1. Score distributions by language group are illustrated in Figure 5.4. Reliability was calculated using the Kuder-Richardson formula 20 and ICC2k, and was generally very high both in L1 (KR-20: 0.95, SEm = 2.32, ICC2k = 0.81 [CI 95% 0.73–0.87]) and L2 (KR-20: 0.93, SEm = 2.37, ICC2k = 0.76 [CI 95% 0.67–0.84]). Although these reliability values

are higher than those calculated for AFT, it is worth reiterating that the two measures are calculated differently and are not directly comparable.

LexTALE accuracy scores for all items (i.e. words and non-words) ranged between 76.67% and 100% in L1 ($M = 93.77\%$, $SD = 5.45$), and 61.67% and 100% in L2 ($M = 81.47\%$, $SD = 10.70$), see Table 5.3.1. Score distributions by language group are illustrated in the raincloud plot in Figure 5.4. Reliability was considerably lower in L1 (KR-20: 0.64, SEm = 1.27, ICC2k = 0.61 [CI 95% 0.50–0.75]) compared to L2 (KR-20: 0.86, SEm = 2.16, ICC2k = 0.83 [CI 95% 0.75–0.89]), reflecting its intended use as a measure of second language proficiency. Correspondingly, LexTALE appears to be less reliable in our L1 sample, and consequently, less informative.

Semantic fluency scores for grocery items ranged between 6 and 28 in L1 ($M = 18.34$, $SD = 4.59$), and 7 and 24 in L2 ($M = 14.77$, $SD = 4.3$), and scores for public figure names ranged between 4 and 19 in L1 ($M = 10.62$, $SD = 3.25$), and 7 and 16 in L2 ($M = 9.91$, $SD = 2.78$), see Table 5.3.1. Score distributions by language group are illustrated in the raincloud plot in Figure 5.4. Test-retest reliability was also calculated for both general semantic fluency tasks. In the L1 sample, 45/50 participants returned, and we calculated an acceptable TRR for the grocery items (Pearson's $r = 0.76$, $p < .001$) and a moderate TRR for public figure fluency (Pearson's $r = 0.67$, $p < .001$).

In L2, just 20/50 of the initial respondents returned to retake the semantic fluency task. Shapiro-Wilk tests revealed the semantic fluency subtasks at time 1 and time 2 were not normally distributed in the L2 sample. For this reason, we

calculated TRR using Spearman's correlation for the grocery items ($r = 0.48$, $p < .05$) and for the public figure fluency subtask, which only approached significance ($r = 0.39$, $p = .09$). These TRR values were significantly lower than those for AFT, although this was likely due in part to the even lower sample size—however, we observe that the TRR values in the larger L1 sample were also somewhat lower for the semantic fluency tasks than those for AFT. Below, unless otherwise stated, we report only on the scores for the repeated tasks (AFT, semantic fluency) at time 1.

Table 5.3.1: Summary statistics for each predictor. Both groups n= 50.

	Group	Min	Median	Mean	Max	SD	IQR	W	*p*-value
AFT	L1	1	10.5	10.16	28	5.75	7.00	6292	**<.01****
	L2	-1	8	7.96	27	5.73	6.75		
ART	L1	4	26	25.62	42	10.27	17.50	6394	**<.001*****
	L2	2	20	20.92	45	9.04	13.25		
LexTALE Accuracy	L1	76.67	95.00	93.77	100	5.45	6.25	8386	**<.001*****
	L2	61.67	81.67	81.47	100	10.7	19.58		
LexTALE Words Accuracy	L1	77.50	95.00	94.95	100	5.33	7.50	8224	**<.001*****
	L2	47.50	81.25	79.30	100	14.53	25.00		
LexTALE Non-Words Accuracy	L1	50.00	95.00	91.40	100	11.07	10.00	6712	**<.001*****
	L2	45.00	85.00	85.80	100	11.92	15.00		
SF Grocery Items	L1	6	19	18.34	28	4.59	6.75	5024	**<.001*****
	L2	7	15	14.77	24	4.30	4.50		
SF Public Figures	L1	4	10	10.62	19	3.25	4.50	3892	.11
	L2	7	9	9.91	16	2.78	4.75		

Note: Wilcoxon tests used due to variables not being normally distributed. LexTALE accuracy scores are in percentages. SF = semantic fluency. Significant correlations are in **bold**; * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

5.3.2 Outcome Variables

Visual world paradigm accuracy scores per group and sentence type are recorded in Table 5.3.2. On idiomatic trials, L1 scores ranged between 84.85% and 100% (M = 97.39, SD = 3.52), and L2 scores were between 60.61% and 100% (M = 93.31, SD = 9.46). On literal trials, L1 scores ranged between 81.82% and 100% (M = 94.55, SD = 5.23), and L2 scores were between 63.64% and 100% (M = 85.35, SD = 10.56). Performance differed significantly between L1 and L2 participants on literal (W = 1942, $p < .001$), but not idiomatic trials (W = 1416, $p = .22$). In both conditions, participants from both language groups generally performed close to ceiling, as illustrated in Figure 5.5. This demonstrates that participants from both groups found it relatively easy to understand which image best corresponded to each sentence. However, we do also note an expected—albeit slight—formulaic advantage for the idiomatic sentences compared to literal trials.

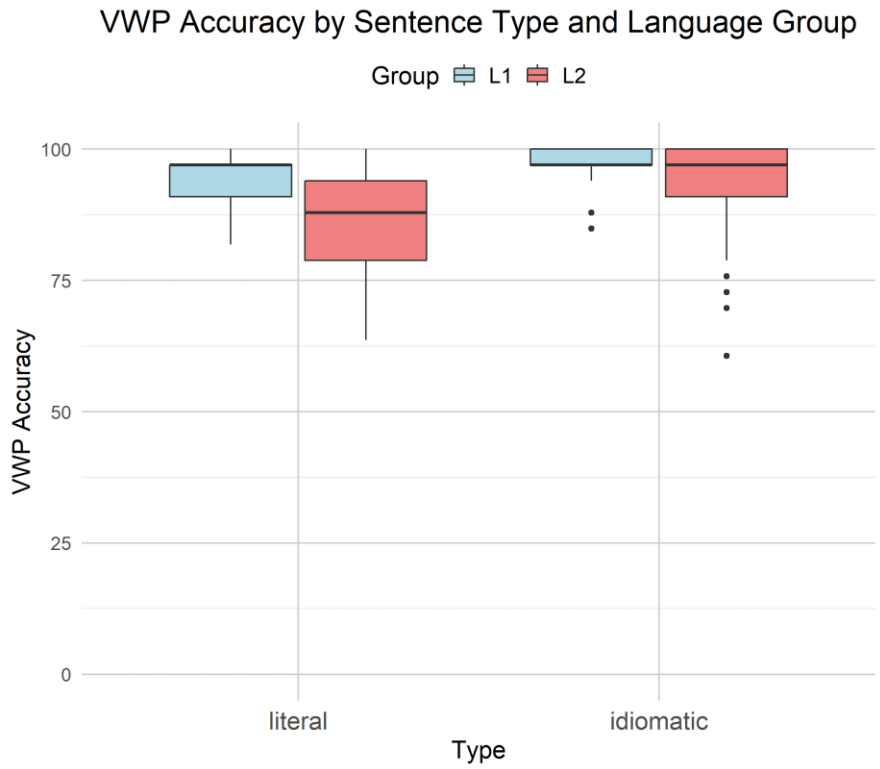


Figure 5.5: Bar plot illustrating image selection accuracy on the visual world paradigm, by language group (L1/L2) and sentence type (literal/idiomatic). Performance was approximately at ceiling in both groups, and in both conditions.

Given that the items were approximately equal in difficulty, and that most participants performed near ceiling at the image selection task, the remaining question of where participants looked prior to the onset of the critical word(s) in each sentence is the more interesting and informative one. Eye gaze was recorded using custom scripting for Gorilla Experiment Builder, which provided x/y coordinates for eye positions over time per trial. Example gaze fixations for all participants for sample trials are shown in Figure 5.6. In this figure, gaze fixations during the prediction window for each sentence are averaged across all participants

in each language group. As image positions varied by participant and trial, the concentrations of looks appear to be mostly random.

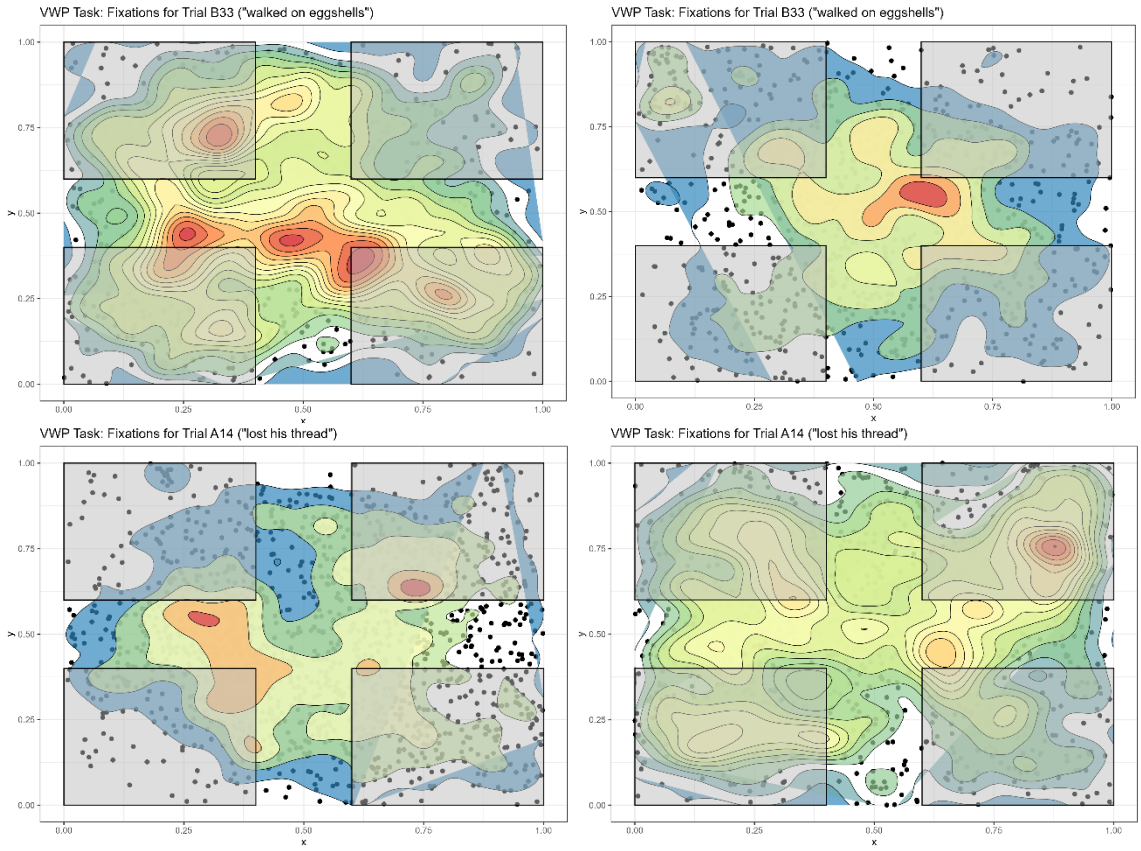


Figure 5.6: Example gaze fixations for various trials for L1 (left) and L2 (right) participants.

Averaged proportions of looks to targets, foils (competitors) and unrelated distractors for idiomatic trials by language group are shown in Figure 5.7. For this and later eye-tracking analyses, we filtered to only those trials in which participants ultimately responded correctly to each image stimulus.

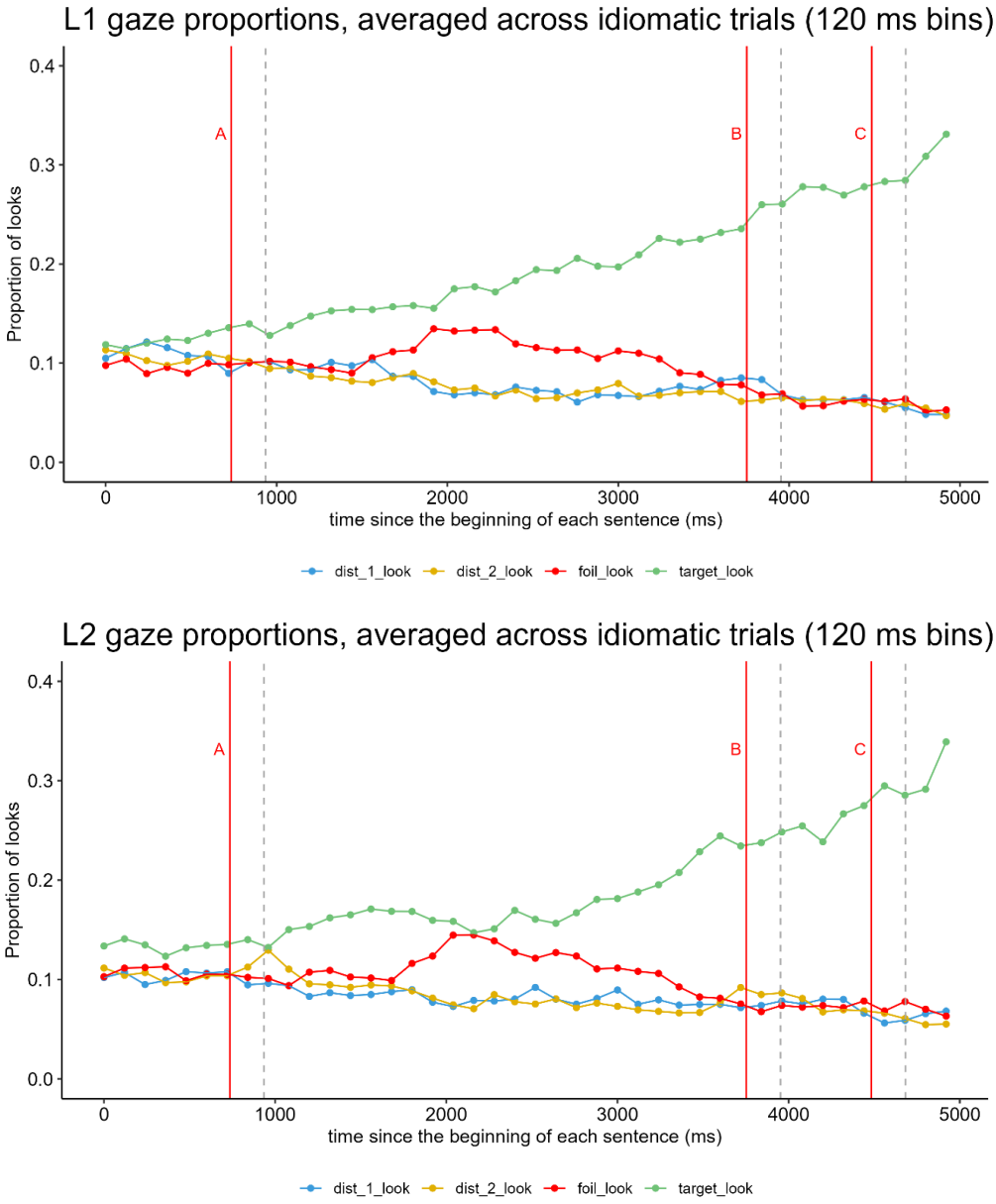


Figure 5.7: Graph showing average proportion of looks to targets (green), foils (red), and distractors (blue and yellow) over time during idiomatic trials, in L1 (top) and L2 (bottom). Binning time was set to 120 ms. Solid red vertical lines indicate regions of interest, with averages for: A) idiom cue onset; B) critical word onset; C) sentence completion. Dashed grey lines indicate +200 ms adjusted time points to account for oculomotor movement latency in response to language processing.

As mentioned previously, we restricted our primary analyses to the "prediction window"; that is, 200 ms after the onset of the cue phrase until 200 ms after the onset of the critical word or phrase. This adjustment is to account for eye movement latency during language processing (Barr, 2008; Matin et al., 1993; Saslow, 1967). Participants' fixation count proportions (per trial and region) were transformed to log odds using the empirical logit function (Barr, 2008; see equation 2).

$$\text{logit}(p) = log(\frac{p}{1-p}) \quad 2$$

This transformation is used because eye movements in a visual world paradigm are by nature non-independent of one another—as Barr (2008) explains, this becomes clear when considering that the eye's position at any given point in time is logically dependent on the position which immediately preceded it. Consequently, the empirical logit function serves to transform the proportional scale into a log odds scale appropriate for estimating multiplicative effects such as those found in eye tracking data; without taking this into account, it is likely that we would overestimate the size and significance of any potential effect. In performing the transformation, we included a constant of 1 (Laplace smoothing) to avoid taking the logarithm of zero in cases where fixation counts were null. Next, following similar methods to those described by Favier et al. (2021), we subtracted the averaged log odds of looks to foils and distractors from the averaged log odds of looks to targets, per participant and per trial, within the prediction window. This difference score, as a de facto measure of target preference, served as the dependent variable for our models. Note that in our analyses, we do not consider the effect of

time as a continuous variable. Instead, following the pre-registered design, we only considered the time windows more broadly (i.e. before and after the beginning of each sentence recording), having anticipated that the webcam-based eye-tracking procedure might not be fast or accurate enough to provide a more fine-grained level of detail.

However, while analysing the time course graphs in Figure 5.7, it appeared as though some participants may have already fixated the target images prior to the average idiomatic cue onset. In other words, it seems that some participants could have been anticipating which image would correspond to a figurative meaning before even hearing the idiomatic phrase. Given that the experimental design was not hidden from participants, this is somewhat unsurprising. Initially, we expected that these "anticipatory baseline effects" (Barr et al., 2011) would be negligible due to the experimental design—essentially, given that there were an equal number of idiomatic and literal trials, prioritising an idiomatic or literal interpretation before hearing the sentence was as likely to be incorrect as correct, meaning it wouldn't necessarily be a productive strategy. Even so, participants may have scrutinised the images for potential idiomatic meanings prior to any additional information. To account for baseline effects statistically, we performed the same target preference calculations and transformations in the time window prior to the onset of the sentence recording, and include this as a covariate in our models. In other words, we controlled for the probability of fixating on a particular image prior to hearing the sentence. While not pre-registered, this decision was taken to account for the possibility that some targets were simply more interesting than others.

Table 5.3.2: Summary statistics for each outcome variable on visual world paradigm task. Each group (English L1 L2) n= 50. Wilcoxon tests used due to variables not being normally distributed. Proportions of looks-to-targets measures looks only within the prediction window. Looks-to-targets percentages calculated by first determining averages per participant by sentence type, then calculating statistics for each sentence type by language group using these averages.

	Group	**Min**	**Median**	**Mean**	**Mode**	**Max**	**SD**	**IQR**	**W**	***p*-value**
Accuracy (%, all trials)	L1	81.82	96.97	95.97	96.97	100	4.66	6.06	6644	**<.001 *****
	L2	60.61	92.42	89.33	100	100	10.74	12.12		
Accuracy (%, idiomatic trials)	L1	84.85	96.97	97.39	100	100	3.52	3.03	5664	.09
	L2	60.61	96.97	93.31	100	100	9.46	9.09		
Accuracy (%, literal trials)	L1	81.82	96.97	94.55	96.97	100	5.23	5.30	7768	**<.001 *****
	L2	63.64	87.88	85.35	90.91	100	10.56	15.15		
Proportion looks-to-target (%, all trials)	L1	-11.39	1.84	1.39	9.11	4.26	3.38	-11.39	5274	.50
	L2	-6.22	1.56	1.94	11.19	3.79	4.11	-6.22		
Proportion looks-to-target (%, idiomatic trials)	L1	-0.72	0.96	1.15	0.57	3.06	0.95	1.57	1432	.21
	L2	-2.02	0.97	0.87	0.20	3.20	1.00	1.14		
Proportion looks-to-target (%, literal trials)	L1	-1.77	-0.05	-0.08	0.12	1.58	0.89	1.30	1238	.94
	L2	-3.13	-0.08	-0.05	0.68	3.31	1.05	1.12		

Table 5.3.3: Correlation matrix for all tasks, L1 participants.

	AFT	*ART*	*LT*	*VWP*	*VWP Idi.*	*VWP Lit.*	*SF: GR.*	*SF: PF*
AFT	1							
ART	**.763 *****	1						
LT	**.416 *****	**.679 *****	1					
VWP	.091	.165	**.417 *****	1				
VWP Idi.	**.345 *****	**.472 *****	**.625 *****	**.453 *****	1			
VWP Lit.	-.069	-.022	**.327 *****	**.610 *****	.139	1		
SF:GR	**.242 ***	.086	.16	**.288 ****	**.309 ****	**.308 ****	1	
SF:PF	**.200 ***	.121	.125	.192	**.220 ***	.196	**.556 *****	1

LT = LexTALE accuracy score for both word and non-word trials. VWP = visual world paradigm accuracy for both idiomatic and literal trials. SF:GR = semantic fluency grocery items. SF:PF = semantic fluency public figure items. Significance codes: ** p<0.05; ** p<0.01; *** p<0.001.*

Table 5.3.4: Correlation matrix for all tasks, L2 participants.

	AFT	*ART*	*LT*	*VWP*	*VWP Idi.*	*VWP Lit*	*SF: GR.*	*SF: PF*
AFT	1							
ART	**.706 *****	1						
LT	**.406 *****	**.540 *****	1					
VWP	**.363 *****	**.342 *****	**.514 *****	1				
VWP Idi.	**.407 *****	**.302 ****	**.458 *****	**.672 *****	1			
VWP Lit	**.380 *****	**.431 *****	**.646 *****	**.695 *****	**.479 *****	1		
SF:GR	**.309 ****	**.303 ***	.163	**.371 ****	**.430 *****	**.374 ****	1	
SF:PF	.144	.041	-.155	.086	.07	.118	**.264 ***	1

LT = LexTALE accuracy score for both word and non-word trials. VWP = visual world paradigm accuracy for both idiomatic and literal trials. SF:GR = semantic fluency grocery items. SF:PF = semantic fluency public figure items. Significance codes: ** p<0.05; ** p<0.01; *** p<0.001.*

5.3.3 H1: Predicting target preference looking from print exposure

Our first hypothesis (H1) was that participants with greater L2 print exposure (as indexed by AFT) would make increased anticipatory looks to target images for idiomatic sentences. To assess this, we built linear mixed-effect models which modelled the increase in target preference in the prediction window as a function

of the fixed effects of print exposure and its interactions with sentence type (literal/idiomatic) and language group (L1/L2), as well as random intercepts for participants and items. As mentioned, baseline target preference was entered as a covariate in the model. All other predictors were first centred and scaled (normalised) prior to entry into the models. On an exploratory basis, we also compared performance between models with ART, either as a covariate or as sole predictor of print exposure.

Comparing otherwise identical models showed a preference for a model with AFT as our predictor of print exposure rather than ART (ΔAIC/BIC = -28.00). In this model (Table 5.3.5), we observed a main effect for baseline target preference ($\beta = 0.17$, $F(1, 5482.8) = 253.65$, $p < .001$), and for sentence type, revealing the idiomatic advantage noted earlier across both language groups ($\beta = 1.11$, $F(1, 119.3) = 58.38$, $p < .001$), and a significant three-way interaction of AFT on sentence type and language group, such that the Author Fluency Task was a significantly higher predictor of target preference for idiomatic trials in the L2 compared to L1 sample ($\beta = 0.92$, $F(1, 5373.6) = 20.84$, $p < .001$). We also evaluated a model with an additional interaction term for baseline target preference on language group and sentence type, but the Bayesian Information Criterion showed a substantial preference (ΔBIC: -85.43) for the simpler model described here. The full model output is provided in Table 5.3.5, and the interaction of AFT, sentence type and language group is illustrated in Figure 5.8.

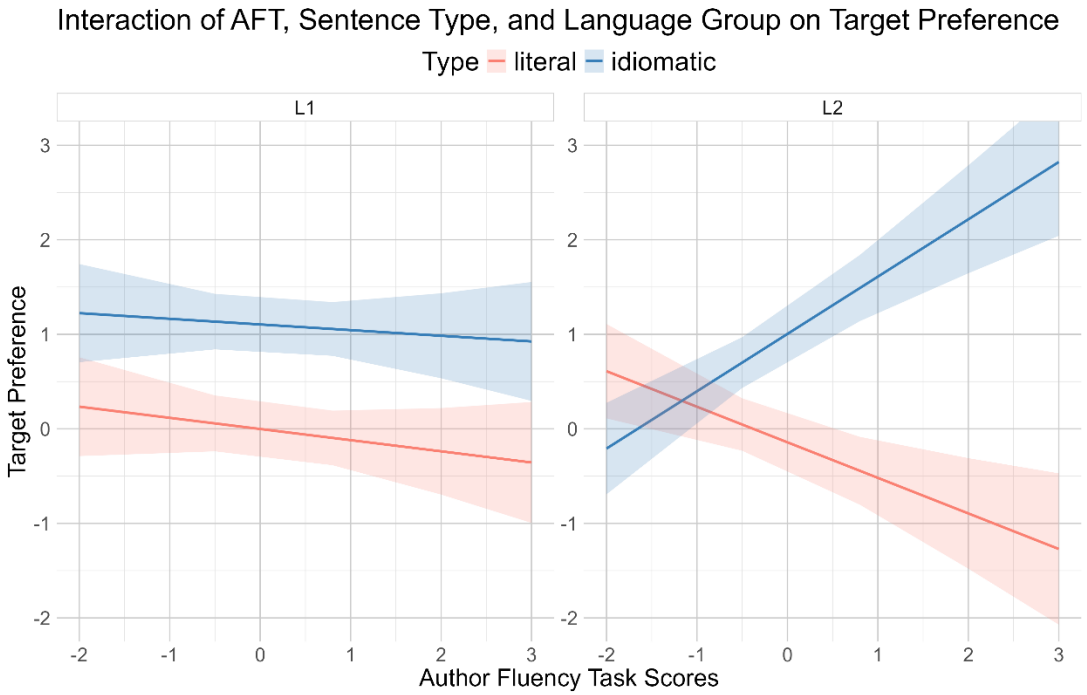


Figure 5.8: Plot showing a three-way interaction between the fixed effects of AFT scores, sentence type, and language group on target preference.

Table 5.3.5: Linear mixed effects model predicting target preference, showing a main effect of sentence type and baseline target preference, and a three-way interaction between AFT, sentence type, and language group.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.009	-0.271 – 0.252	0.945
Baseline Target Preference	0.172	0.151 – 0.193	**<0.001*** **
AFT	-0.118	-0.326 – 0.090	0.265
Type [Idiomatic]	1.106	0.763 – 1.449	**<0.001*** **
Group [L2]	-0.141	-0.454 – 0.173	0.379
AFT × Type [Idiomatic]	0.058	-0.204 – 0.320	0.664

AFT × Group [L2]	-0.258	-0.572 – 0.055	0.107
Type [Idiomatic] * Group	0.040	-0.354 – 0.435	0.841
(AFT × Type [Idiomatic]) × Group [L2]	0.924	0.527 – 1.321	**<0.001*** **
Random Effects			
σ^2	13.25		
$\tau_{00\ Item}$	0.35		
$\tau_{00\ Participant}$	0.11		
ICC	0.03		
$N_{\ Participant}$	100		
$N_{\ Item}$	120		
Observations	5526		
Marginal R^2 / Conditional R^2	0.072 / 0.103		

** p<0.05 ** p<0.01 *** p<0.001*

Clearly, whatever aspect of print exposure AFT indexes in L2, the same cannot be said for the L1 sample. In comparison, an analogous model with ART as the sole predictor of print exposure (Table 5.3.6) shows that ART is a non-significant predictor of target preference, whereas we see virtually identical main effects for both baseline target preference ($\beta = 0.17$, $F(1, 5482.7) = 253.79$, $p < .001$), and for sentence type ($\beta = 1.07$, $F(1, 121.7) = 52.48$, $p < .001$) as those found in the previous model (Table 5.3.5).

Table 5.3.6: Linear mixed effects model predicting target preference, showing a main effect of sentence type and baseline target preference. The three-way interaction between ART, sentence type, and language group is non-significant.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.001	-0.264 – 0.262	0.995
Baseline Target Preference	0.173	0.151 – 0.194	**<0.001*** **
ART	-0.135	-0.338 – 0.069	0.194
Type [Idiomatic]	1.066	0.720 – 1.411	**<0.001*** **
Group [L2]	-0.106	-0.423 – 0.211	0.511
ART × [Idiomatic]	0.216	-0.041 – 0.472	0.100
ART × Group [L2]	-0.011	-0.336 – 0.314	0.946
Type [Idiomatic] × Group [L2]	0.016	-0.383 – 0.416	0.936
(ART × Type [Idiomatic]) × Group [L2]	0.339	-0.070 – 0.748	0.104
Random Effects			
σ^2	13.32		
$\tau_{00\ Item}$	0.35		
$\tau_{00\ Participant}$	0.11		
ICC	0.03		
$N_{\ Participant}$	100		
$N_{\ Item}$	120		
Observations	5526		
Marginal R^2 / Conditional R^2	0.067 / 0.098		

** p<0.05 ** p<0.01 *** p<0.001*

The distinction between the two print exposure measures becomes clearer when evaluating a model which includes both, along with their interactions on language group and sentence type, and random intercepts for participants and items. We observed a main effect for trial type, revealing main effects for baseline target preference ($\beta = 0.17$, $F(1, 5475.5) = 251.49$, $p < .001$), and the idiomatic sentence advantage ($\beta = 1.07$, $F(1, 121.9) = 54.71$, $p < .001$), as well as a positive interaction for ART and sentence type ($\beta = 0.41$, $F(1, 5334.2) = 0.30$, $p = .04$), but this was significantly moderated by the negative effect in the L2 group ($\beta = -0.65$, $F(1, 5362.5) = 4.88$, $p = .03$)—which not only cancels out the effect of ART on sentence type in L2, but in fact reverses it. A significantly negative interaction of AFT on language group ($\beta = -0.48$, $F(1, 96.3) = 1.55$, $p = .04$) was also reversed by the significant and positive three-way interaction of AFT on sentence type and language group, such that AFT was a significantly greater predictor of target preference for idiomatic trials in L2 compared to L1 ($\beta = 1.39$, $F(1, 5330.5) = 23.20$, $p < .001$). The full model output is shown in Table 5.3.7, and an illustration comparing effects of AFT and ART by sentence type and language group are shown in Figure 5.9.

Table 5.3.7: Linear mixed effects model predicting target preference, showing a main effect of sentence type and baseline target preference, and a three-way positive interaction between AFT, sentence type, and language group. The same interaction with ART was significantly negative. A positive interaction between ART and sentence type was reversed in L2 when controlling for AFT.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.001	-0.264 – 0.262	0.996
Baseline Target Preference	0.172	0.150 – 0.193	**<0.001*** **
AFT	-0.032	-0.353 – 0.289	0.846
ART	-0.111	-0.425 – 0.202	0.486
Type [Idiomatic]	1.072	0.727 – 1.417	**<0.001*** **
Group [L2]	-0.125	-0.443 – 0.193	0.441
AFT × Type [Idiomatic]	-0.258	-0.660 – 0.145	0.210
ART × Type [Idiomatic]	0.407	0.014 – 0.801	**0.042***
AFT × Group [L2]	-0.477	-0.927 – -0.028	**0.037***
ART × Group [L2]	0.328	-0.134 – 0.790	0.164
Type [Idiomatic] × Group [L2]	0.047	-0.352 – 0.446	0.817
(AFT × Type [Idiomatic]) × Group [L2]	1.392	0.825 – 1.958	**<0.001*** **
(ART × Type [Idiomatic]) × Group [L2]	-0.652	-1.231 – -0.074	**0.027***
Random Effects			
σ^2	13.24		
$\tau_{00\ Item}$	0.35		

$\tau_{00\ Participant}$	0.11
ICC	0.03
$N_{\ Participant}$	100
$N_{\ Item}$	120
Observations	5526
Marginal R^2 / Conditional R^2	0.073 / 0.104

** p<0.05 ** p<0.01 *** p<0.001*

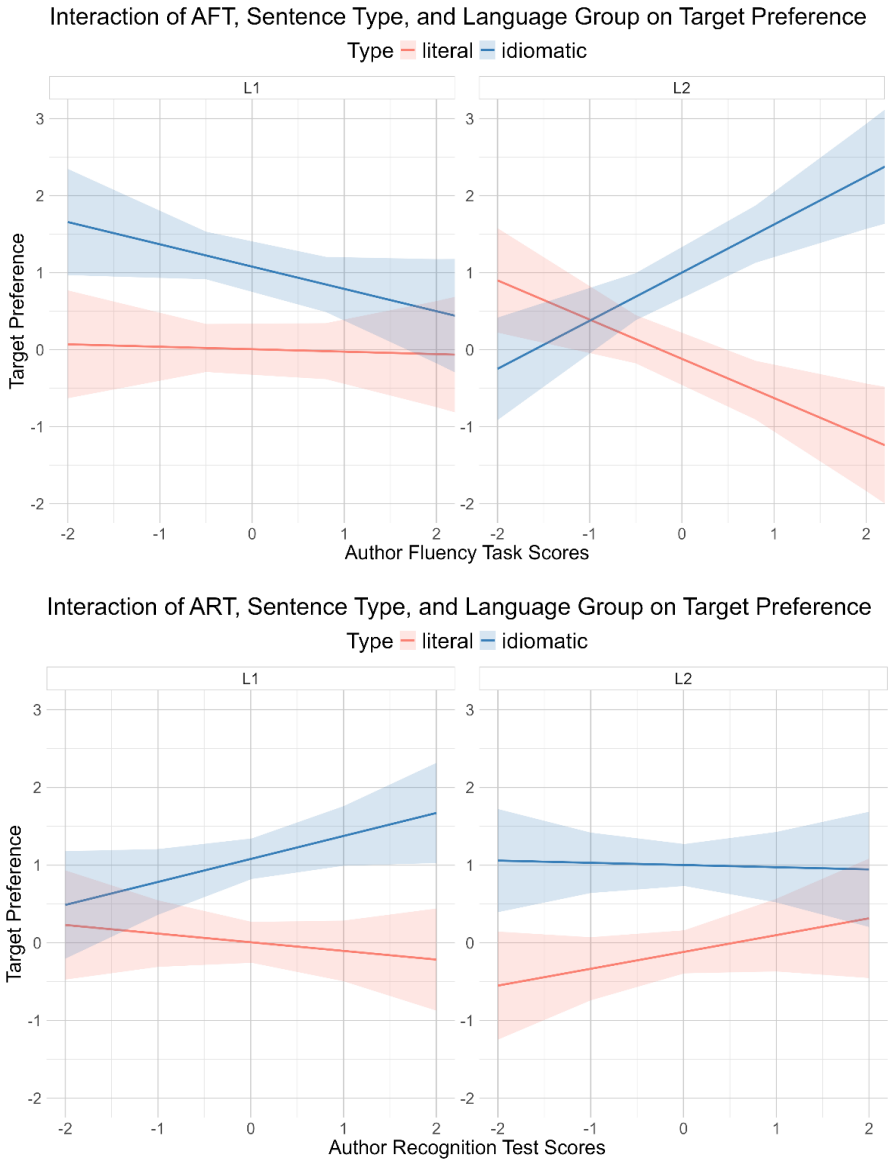


Figure 5.9: Effects of AFT (top) and ART (bottom) scores on proportions of looks-to-targets in the prediction window prior to the onset of critical word(s), by language group (L1/L2) and sentence type (literal/idiomatic). These effects are from a model where both print exposure measures (AFT and ART) were included; as a result, the meagre effect of ART in L2 is likely primarily a reflection of the limited additional predictive value of ART once controlling for AFT in this population.

On an exploratory basis, we also evaluated the contributions of verbal fluency skills more generally to predictive target preferences, and to determine if this would eliminate the effect of AFT. To do so, we added scores for both subtasks of the semantic fluency task (Grocery Items, Public Figures) to the initial model in Table 5.3.6. In this expanded model, we again observed a main effect for baseline target preference ($\beta = 0.18$, $F(1, 4648.7) = 235.84$, $p < .001$), and for sentence type, reflecting the idiomatic advantage ($\beta = 1.17$, $F(1, 152.6) = 54.35$, $p < .001$), and a significant three-way interaction of AFT on sentence type and language group, such that the Author Fluency Task was a significantly higher predictor of target preference for idiomatic trials in the L2 compared to L1 sample ($\beta = 0.62$, $F(1, 4515.2) = 6.58$, $p < .05$). These lower estimates for AFT compared to the model in Table 5.3.6 appear to be due to the shared variance with the scores from the grocery items subtask scores, which also showed a significant positive interaction with group and sentence type ($\beta = 0.63$, $F(1, 4569.7) = 5.60$, $p < .05$). Interestingly, the main effect of public figure fluency scores, as well as their interaction with language group and sentence type, were all non-significant. The same general pattern of results was found when evaluating models where each semantic fluency subtask was entered in separately as a covariate with AFT, suggesting the patterns described here were not simply due to the fluency tasks being correlated. The full output for this model is shown in Table 5.3.8.

Table 5.3.8: Linear mixed effects model predicting target preference, showing a main effect of sentence type and baseline target preference, and a three-way positive interaction between AFT, sentence type, and language group. The same interaction with semantic fluency grocery items was also positive, whereas semantic fluency public figure items were non-significant.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.012	-0.273 – 0.249	0.927
Baseline Target Preference	0.178	0.155 – 0.200	**<0.001*** **
AFT	-0.125	-0.335 – 0.086	0.247
SF Grocery	0.020	-0.245 – 0.284	0.885
SF Public Figures	0.016	-0.224 – 0.257	0.896
Type [Idiomatic]	1.174	0.832 – 1.516	**<0.001*** **
Group [L2]	0.008	-0.363 – 0.379	0.967
AFT × Type [Idiomatic]	0.084	-0.184 – 0.352	0.540
SF Grocery × Type [Idiomatic]	-0.335	-0.670 – 0.001	0.050
SF Public Figures × Type [Idiomatic]	0.238	-0.067 – 0.542	0.126
AFT × Group [L2]	-0.004	-0.377 – 0.369	0.985
SF Grocery × Group [L2]	-0.063	-0.476 – 0.350	0.766
SF Public Figures × Group [L2]	-0.119	-0.508 – 0.269	0.548
Type [Idiomatic] × Group [L2]	-0.043	-0.513 – 0.427	0.858
(AFT × Type [Idiomatic]) × Group [L2]	0.622	0.147 – 1.097	**0.010***

(Grocery × Type [Idiomatic]) × Group [L2]	0.630	0.108 – 1.151	**0.018***
(Public Figures × Type [Idiomatic]) × Group [L2]	-0.218	-0.707 – 0.271	0.383
Random Effects			
σ^2	12.97		
$\tau_{00\ Item}$	0.30		
$\tau_{00\ Participant}$	0.10		
ICC	0.03		
$N_{\ Participant}$	84		
$N_{\ Item}$	120		
Observations	4693		
Marginal R^2 / Conditional R^2	0.074 / 0.102		

** p<0.05 ** p<0.01 *** p<0.001*

5.3.4 H2: Effects of print exposure on VWP accuracy

Our second hypothesis (H2) was that L2 print exposure would correlate positively with image selection accuracy on the VWP task. To assess this, our preregistration initially specified that we would use multiple regression models. However, we ultimately realised that such models would not be appropriate for these analyses, because we would not be able to examine trial-level data without violating the assumption of independence of observations. Instead, we decided to use mixed-

effects models which do not have this same requirement, and which would also permit us to estimate random effects for items and participants. Thus we built generalised mixed-effects models predicting the relative odds of correct responses on each trial of the VWP selection task. Contrasts were treatment (dummy) coded, with the baseline set to "literal" for sentence type.

We observed a significant main effect of AFT, (OR = 1.07, p < .01) corresponding to an increase of 7% in odds of correct selections for each individual author named. There was also a significant interaction between AFT and sentence type, such that AFT predicted increased odds ratios for idiomatic sentences (OR = 1.12, p < .001), or 12% for each author named. For reference, in an analogous model with ART as the measure of print exposure, an increase in one point on ART was associated with a 4% increase in VWP accuracy, and with an additional 4% increase in odds ratios for idiomatic sentences, i.e. ⅓ of the corresponding increase using AFT scores. The AFT model was also substantially preferred to the ART model (ΔAIC/BIC = -10.25).

Although not mentioned in the preregistration, we considered that it may also be informative to compare VWP performance as a function of print exposure between L1 and L2 groups. This final exploratory model included fixed effects of AFT and its interaction with sentence type and language group, as well as random intercepts for participants and items (Marginal-R^2: .17, Conditional-R^2: .50; Table 5.3.9). As before, a model with AFT as the print exposure measure was substantially preferred to an ART model (ΔAIC/BIC = -13.57). Contrasts were treatment (dummy) coded, with the baseline set to "literal" for sentence type, and "L1" for group. We observed a significant main effect of language group, such that

L2 speakers were 86% less likely to make correct selections on the VWP task compared to L1 speakers, across both sentence types ($OR = 0.14$, $p < .001$). The main effect for AFT was non-significant, but there was a significant interaction between AFT and language group, such that AFT predicted slightly increased odds ratios in L2 ($OR = 1.10$, $p < .05$). In other words, for each additional author named by L2 speakers, the odds of correct selections increased by 10% compared to L1 speakers. This suggests that higher AFT scores may partially mitigate the disadvantages associated with being an L2 speaker. To illustrate this point, assuming this 10% odds increase should continue linearly, an L2 speaker who names approximately nine additional authors would be predicted to offset this negative group effect. Finally, there was a significant interaction between AFT scores and idiomatic sentence trials ($OR = 1.13$, $p < .01$), such that each additional author name increased the odds of correct selections for idiomatic trials by 13% across both L1 and L2 groups. This suggests that print exposure is particularly associated with knowledge of formulaic rather than literal language. Fixed effects and interactions are visualised in the forest plot in Figure 5.10.

Table 5.3.9: Mixed effects model predicting odds of correct VWP image selections, showing fixed effects of AFT and interactions with language group and sentence type, and random effects of participant/item.

	Correct		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	44.753	20.869 – 95.970	**<0.001*** **
AFT	0.990	0.934 – 1.050	0.737
Type [Idiomatic]	0.744	0.296 – 1.870	0.530
Group [L2]	0.139	0.060 – 0.321	**<0.001*** **
AFT × Type [Idiomatic]	1.132	1.042 – 1.228	**0.003****
AFT × Group [L2]	1.099	1.016 – 1.190	**0.019***
Type [Idiomatic] × Group [L2]	1.936	0.764 – 4.905	0.164
(AFT × Type [Idiomatic]) × Group [L2]	1.009	0.905 – 1.125	0.875
Random Effects			
σ^2	3.29		
$\tau_{00\ Item}$	1.50		
$\tau_{00\ Participant}$	0.67		
ICC	0.40		
$N_{\ Participant}$	101		
$N_{\ Item}$	132		
Observations	6666		
Marginal R^2 / Conditional R^2	0.173 / 0.502		

** p<0.05 ** p<0.01 *** p<0.001*

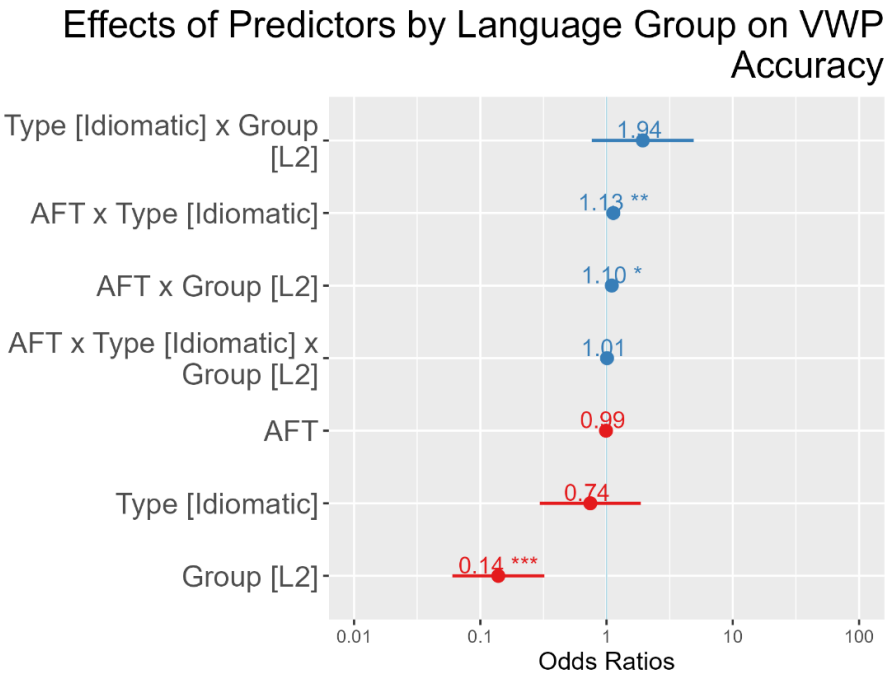


Figure 5.10: Fixed effects with interactions for language group and sentence type, predicting odds of correct image selections on VWP task. Lines represent the 95% confidence interval.

5.4 General Discussion

This study served two main theoretical and methodological aims. With regards to theory, we examined how print exposure is associated with online, predictive processing and offline, explicit knowledge of idioms in L2. Methodologically, we evaluated the use of relatively new webcam eye-tracking technology for visual world paradigm studies. Results demonstrated that although both L1 and L2 speakers accurately understood the meanings of idioms in the offline judgement task and performed equivalently, we found that print exposure was significantly associated

with target preference for idiomatic trials in L2 speakers. This finding confirmed H1, and show that reading experience is correlated with early online predictive processing. Moreover, these results are consistent with the idea that experience with written language transfers to the spoken domain in L2 speakers, just as it has been shown to do previously in L1 speakers. Additionally, although accuracy scores were near ceiling for both literal and idiomatic trials and across language groups, results showed that L2 participants with greater print exposure were significantly more likely to accurately select images in both conditions, confirming H2.

Further scrutiny provides some indication of the strategies used by L1 and L2 speakers. As Figure 5.8 illustrates, as print exposure increases, target preference tends to increase commensurately in L2, but not L1 speakers. Interestingly however, target preference also *decreases* as a function of print exposure in the literal sentence condition for L2, but not L1 speakers. In other words, it appears that L2 speakers with more reading experience tend to avoid settling for a straightforward, literal interpretation in literal trials, and instead continue searching for a potential idiomatic meaning throughout the sentence. Yet when they encounter a known idiomatic phrase, these L2 speakers with greater reading experience tend to fixate and perseverate on the recognised target. Additionally, L2 participants with the lowest print exposure appear to have a slight preference for the literal interpretation throughout the prediction window of idiomatic sentence trials. This suggests that those with the least L2 reading experience have little familiarity with the idiomatic phrases, and instead are more likely to interpret the sentences literally, using the visual information available. Even so, most L2 speakers ultimately responded correctly to these idiomatic trials after hearing the critical word (M = 93.31%,

mode = 100%). This indicates that although they generally understood the sentence meaning in the offline comprehension task, L2 speakers with varying levels of print exposure were most clearly distinguished during online processing. In essence, these findings show that more experienced L2 readers have earlier and more reliable access to idiomatic sentence meaning.

Exploratory analyses comparing the two measures of print exposure (AFT and ART) also revealed the same distinction between the performance of L1 and L2 speakers as seen in Chapter 3 (McCarron et al., 2025). That is, AFT was more positively associated with outcome measures in L2 than in L1, where ART was a more useful predictor. We found that even when controlling for ART, AFT scores are significantly more associated with preferential looks to targets in the L2 sample (see Figure 5.9). This reinforces the suggestion, established in previous chapters, that ART scores may not be as indicative of L2 reading experience as they are in L1.

Turning to methodological implications, one aspect to consider is the variability in precision and refresh rates in consumer webcams compared to dedicated eye-tracking devices typically found in university laboratories. Sampling rates (measured in Hertz, or the number of samples per second) for lab-based eye trackers are typically between 25 and 2000 Hz (R. Andersson et al., 2010), with most contemporary devices in the higher ends of this scale. In comparison, the sampling rate for webcam eye tracking is constrained by both the webcam and a monitor's refresh rate, which is typically just 60 Hz (Gorilla Experiment Builder, n.d.). Moreover, web-based eye tracking may suffer from variable latency of up to 300 ms (Slim & Hartsuiker, 2023). Naturally, these latency times can vary depending on

an individual's Internet speed and the processing power of the computer being used, and this variation can make it challenging to ensure good temporal resolution, as well as consistent measurement across participants. Finally, whereas lab-based eye trackers can provide information about track loss, eye blinks, and saccades, this information is not available through the JavaScript functions used on Gorilla. Despite these considerations, as Slim and Hartsuiker (2023) point out, webcam eye-tracking can still be useful given the ease of online recruitment, and as a result it may be optimal when temporal latency is less important to the experimental design. This was the case in our experiment, as we only looked at the proportion of looks to targets within a large predictive window, as opposed to a continuous temporal analysis. Undoubtedly, some of the more granular dynamics will be missed with this approach, but we argue these were not essential to our study. To the best of our ability, we ensured that the only indicator of the idiomatic meanings of each sentence was the idiom itself, and the remainder of each sentence was essentially filler to allow participants to anticipate the relevant image. Moreover, the presence of a significant effect of print exposure on anticipatory target preference, in spite of this wide predictive window, suggests that this is likely a true effect. To this point, although the effect sizes for predictive looking behaviour are relatively low ($R^2 \approx 0.10$), evidence suggests that effect sizes may be twice as high in a laboratory eye-tracking setting compared to online studies such as these (Prystauka et al., 2023). Finally, it has also been argued that some statistical analyses may systematically underestimate effect sizes in L2 (Plonsky & Oswald, 2014). Taken together, these factors suggest that our findings are likely indicative of true effects, but would be worth replicating with in-person laboratory studies where temporal

and spatial precision and resolution are much higher. However, because the visual world paradigm only uses the relatively limited four quadrants of the screen to measure eye gaze (compared to a reading study with sentences or paragraphs, in which targets are far more crowded), we suspect the comparatively lower spatial resolution of a consumer webcam was unlikely to be an issue for the current design. Moreover, with regards to temporal resolution, if it were to be the case that the exact timing of looks was not accurately captured, we argue that the within participant design of the study (in which each participant must complete either an idiomatic or literal version of each sentence, but never both) provides some assurance that the overall pattern of data is likely to be reliable.

Experimental psychology studies have increasingly begun taking place online due to the ease of recruitment, but face their own challenges due to problems of variable device latency, participant inattentiveness or poor comprehension of instructions, and selective attrition, among others (Rodd, 2024). Admittedly, attrition was a potential factor in this study—29 L1 participants (16 due to failing device screening, 13 from non-completion, i.e. "time-outs") and 16 L2 participants (13 device screening, 3 non-completion) who signed up for the study were unable to finish, and it is unclear if there were any systematic reasons for these exclusions. Consequently, adapting this task to a lab-based study would be welcome, as these issues could perhaps be mitigated.

Another consideration for this kind of study relates to anticipatory baseline effects (Barr et al., 2011) which we attempted to control for statistically. To do so, we included the baseline target looking preference (prior to hearing the sentence) as a covariate in the models. However, it is also possible that participants were

simply too aware of the experimental design. As discussed previously, we did not attempt to hide the fact that the task related to idiomatic and literal meanings of sentences. Although this was not explained explicitly, it was demonstrated during the practice tasks, which showed both idiomatic and literal versions of three different sentences, which would allow participants to compare both. This introduction to the task may have encouraged participants to seek out images with possible idiomatic associations in the main trials. Of course, half of all sentence stimuli were literal trials, which would presumably dissuade participants from employing strategies like these. However, if the study were to be replicated, it may be worth considering removing these practice sessions altogether, and instead have participants begin with a series of "throwaway" trials, allowing them the opportunity to practice the task without being aware of what exactly is being measured. Alternatively, filler trials with no possible idiomatic meaning could also be included, although it would come at the cost of a longer task overall. Finally, the 2-second "preview" window might provide participants with too much time to consider the images and to make predictions, and it may be ideal to reduce this window to some extent.

One potential objection to the present findings is that perhaps what AFT really measures is verbal fluency skill. Undoubtedly, this is part of the picture. By definition, it is true that an individual's score on any measure of verbal fluency will depend on how many items they can provide within the time limit. In other words, one's score is necessarily constrained by speed and flexibility of lexical access and retrieval of semantically related items from memory. As in the previous study detailed in Chapter 3 (McCarron et al., 2025), we attempted to isolate the

contribution of verbal fluency more generally by including additional measures of semantic fluency, using both common nouns (grocery items) and proper names (public figures). We found that the effect of AFT was only partially mitigated by the common noun fluency scores, but not proper name fluency scores. Naturally, the two proper name measures (authors and public figures) will likely share more variance because they share lexical access pathways, whereas common nouns are retrieved using a different route (Proverbio et al., 2001). Yet within the category of proper names, only knowledge of author names predicts an increased anticipatory target preference, indicating that reading experience is of particular importance for L2 speakers in this regard.

A unique aspect of this study is the use of images as opposed to words for representing idioms. Of course, as the name suggests, images are typically used in visual world paradigms, but we are not aware of any other studies in which idiomatic and literal phrases have been represented visually in this way. At first glance, however, it may not be clear what benefit there is to presenting idioms visually rather than with words. To illustrate, returning to the "kicked the bucket" example, is there any advantage to using images of both a bucket and a gravestone (as distractor and target respectively), rather than showing the words "bucket" in one quadrant, and the final word of the sentence, "dead" in another? In this way, perhaps we could just as easily determine if participants anticipate the idiomatic meaning of the sentence by looking more towards the word "dead" prior to hearing it. This approach has been used effectively in other similar studies (e.g., Kessler & Beck, 2022), and certainly it would have saved us the considerable trouble of having

to create so many different images. But what is gained by making the task primarily visual, as opposed to using written language?

There are a few potential responses to this question, but suffice it to say that image processing is—at least ostensibly—unrelated to language skill, in that the capacity to interpret and understand the semantic contexts that each image represents does not require a certain level of language proficiency. To the best of our ability, we tried to ensure that the iconic images used were broadly universal in their intended meanings.[24] Participants must also maintain the words which they have heard in working memory as they interpret the rest of each sentence, and seeing the words on screen may interfere with this processing. For instance, a participant may see an unrelated distractor word on screen, and misremember if they heard this same word or a phonologically similar one, causing them to fixate on it longer. Essentially, we hoped that participants would primarily focus on processing each sentence, while allowing their eyes to wander in the direction of the most appropriate image, without thinking too much about why. To this point, as the old adage (or cliché) goes, "a picture is worth a thousand words"—that is, images can evoke strong sentiments and contexts using fewer strokes than would be required to write a handful of characters. As a result, it is possible to create vastly different contextual meanings from images with slight modifications to facial expressions, hand gestures, the relative positions of characters and objects, and more.

[24] Nevertheless, we recognise that images and iconography are not universal, and certain images would require adaptation depending on the target population to ensure that they are equally comprehensible. One example is the gravestone featuring a Christian cross, which would not be as familiar in many cultures as a representation of death.

On a final theoretical point, we note that our print exposure measures were associated with both figurative and literal language knowledge in L2, but only with figurative language in L1 (see correlation matrices, L1: Table 5.3.3, L2: Table 5.3.4). Not only this, but for L2 speakers, AFT scores were more highly associated with idiomatic trial accuracy, whereas ART was more associated with literal trial accuracy. If is true that AFT reflects primary print exposure, this finding may reinforce the notion that these more personal experiences with reading help to acquire figurative language. Yet for L1 speakers, this relationship was once again reversed; ART was more strongly associated with idiom accuracy compared to AFT, although both were significantly positive. Once again, we suspect that because ART has been carefully attuned to the reading habits of native speakers, it is more strongly associated with knowledge of formulaic language in this population. In contrast, the meaning of literal language is generally clear enough that print exposure is not a "required ingredient" for its interpretation.

Additionally, we observed that scores for AFT and the "public figure" fluency task were not correlated for L2 speakers, but they were in L1. Again, if AFT scores were to be explained away by a verbal fluency skill more generally, we might expect these scores to correlate across language groups—indeed, we might expect this relationship to be stronger in L2, where fluency is more variable. Moreover, this finding could suggest that in an L1 context, proxy measures of print exposure have more to do with general cultural knowledge. Originally, we suspected that this relationship between cultural knowledge and print exposure was unique to ART in L1, as a symptom of the broader recognition skill required. Instead, it is likely that all such measures—including open-ended ones such as AFT—tend to correlate with

cultural exposure. In this view, through regular exposure to the language, L1 speakers increase their general cultural knowledge, which encompasses both (for example) public figures and authors. As a consequence, this familiarity with authors in L1 may not necessarily be due to personal reading experience. In contrast, L2 speakers receive less cultural exposure overall, making reading of particular importance for developing their knowledge of idiomatic language.

5.5 Data Availability

Data and code used in analyses are available on OSF (https://osf.io/rftnb).

6. General Discussion

This chapter will begin by summarising the current literature landscape on print exposure and language development, particularly in L2, exploring the motivations behind the present research. Following this, I will describe the main findings and theoretical implications of the thesis, arguing that second language print exposure is a critical aspect of developing advanced L2 vocabulary knowledge. Moreover, although the Author Fluency Task does not fundamentally alter our understanding of the importance of print exposure in L2, it does show that the choice of measure may result in different outcomes depending on the population under study. I will also discuss several methodological considerations and limitations, including the challenges of creating a novel visual world paradigm to represent formulaic language.

In the final section, I present some general reflections as an afterword, considering the role of reading experience for developing second language proficiency. Here, I begin by addressing the use of generative AI technologies for producing images for psycholinguistic experiments, as well as the nascent development of large language model chatbots as language tutors which may have significant implications for language learning. I also argue that second language

speakers are not defective native speakers, but rather that proficiency is dependent on experience and exposure, in L2 as in L1. Moreover, because so much of the meaning of language is non-compositional or metaphorical, language requires the creative capacity to make inferences about other minds as well as other possible worlds. For this reason, I suggest that book reading plays an important role in developing these intuitions about the internal lives of others, and that our language in practice necessarily reflects our language experience.

6.1 Background and motivation

Print exposure predicts individual differences in reading and language in L1 (Mol & Bus, 2011; Moore & Gordon, 2015; Stanovich & West, 1989) and has been argued to play a role in L2 as well (Kim & Krashen, 1998; Yamashita, 2013), although its contribution is considerably less clear. Yet the amount an individual reads for pleasure is almost necessarily a function of the pleasure derived from reading, forming a kind of tautological reasoning. This leads to the "Matthew effect": a growing gap between the rich and poor of reading skill (Cunningham & Stanovich, 1998; Mol & Bus, 2011; Stanovich, 1986). Essentially, more gifted readers may find reading enjoyable and rewarding, reaping the benefits of increased exposure. Less proficient readers, however, may find reading more frustrating than gratifying, and avoid picking up books in their free time. As a consequence, their skills may stagnate, making reading even less enjoyable. Reading difficulties may be further compounded in L2, as learners may struggle with more obscure vocabulary than they are likely to encounter in daily speech. Surmounting this difficulty is crucial

for second language acquisition however, because a significant portion of L2 vocabulary is acquired incidentally through reading (Restrepo-Ramos, 2015). Naturally, learners require significant language exposure to reach their full potential in L2. For these reasons, researchers require precise psychometric instruments to quantify L2 speakers' exposure to print.

The Author Recognition Test (ART; Stanovich & West, 1989) is the standard test of print exposure, tasking participants with selecting authors from a checklist of names. As a proxy measure for reading experience, ART is well-validated in L1 populations as a predictor of individual differences in reading skills such as vocabulary (Dąbrowska, 2018), word recognition (Chateau & Jared, 2000), spelling (Stanovich & West, 1989), reading frequency (Acheson et al., 2008; Moore & Gordon, 2015), sentence processing (Acheson et al., 2008), oral language skills (Acheson et al., 2008; Mol & Bus, 2011), reading comprehension, and academic achievement (Mol & Bus, 2011).

The primary advantage of proxy measures of print exposure such as ART is they avoid potential social desirability biases associated with self-report measures like reading surveys (West et al., 1993). Yet a standard ART does not indicate whether recognising an author's name represents personal reading *experience* or general reading *exposure* (*"primary vs. secondary print knowledge"*; Martin-Chang & Gould, 2008). Some research even suggests that ART reflects general cultural knowledge rather than reading-specific experience (Moore & Gordon, 2015; Vermeiren et al., 2022). Ideally, researchers would use a test which measures the latter rather than the former, to the extent that these concepts can be extricated. Despite its widespread use, ART faces concerns about its reliability and validity in

L2 populations (McCarron & Kuperman, 2021; Vermeiren & Brysbaert, 2023). Essentially, L2 speakers generally know very few authors on ART—the question is whether they do not read enough in L2, or if they are simply not reading these particular authors.

L2 speakers naturally have very different cultural exposure to their target language compared to L1 populations. If author names are the fundamental unit of measurement in ART, thought to index the latent trait of print exposure—but cultural exposure is more variable in L2—then a rational next step might be to simply ask participants to name L2 authors. Such measures of semantic fluency require participants to list as many items as possible from a given category in a set time, with one point for each unique and valid item. Surprisingly, this kind of approach has not been formally evaluated in studies of print exposure in the past—or at the very least, these studies have apparently not been published. This may be because ART serves its purpose well in L1 populations, and because there has only recently been discussion surrounding the appropriateness of ART in L2 populations. Throughout this thesis, a novel fluency measure of print exposure, the Author Fluency Task (AFT), was the main measure of L2 reading experience used.

6.2 Summary of Theoretical and Methodological Contributions

6.2.1 Summary of findings

In the first experimental study described in Chapter 2, we asked whether a fluency task for author names such as AFT could be used as an alternative to ART for measuring print exposure in an L1 English, university-educated population. We used two main outcome variables and examined how the AFT and ART differed in terms of their predictions. Compared to ART, AFT scores were more highly associated with scores on a self-report reading habits survey. The second outcome variable was selection accuracy on a novel lexical decision task. This used lower-prevalence keywords of literary fiction (compared to non-fiction keywords, obtained from a large corpus of books of various genres). Results showed that compared to AFT, ART was more highly correlated with lexical decision performance. However, an exploratory measure which combined author names from both ART and AFT performed better than either individual measure alone, suggesting that the two proxy measures of print exposure might be distinguishable in some ways.

The main populations of interest of this thesis, however, were proficient L2 speakers of English; specifically, L1 French and Korean speakers of English, who were tested on L2 formulaic language in Chapters 3 and 4. These groups were selected for their similarity and dissimilarity to English respectively, both in terms of language and culture. Comparing AFT and ART for L1 Korean speakers, the

two measures of print exposure were almost equally correlated with accuracy scores for both connectives (e.g., "nevertheless", "indeed", "given that") and collocations (e.g., "raise prices" as opposed to "lift prices"), suggesting either measure of print exposure may be suitable in this population. In the L1 French sample however, we found that compared to ART, AFT was more strongly associated with both connectives and collocations scores in L2, even when controlling for L2 proficiency using an English word recognition measure (LexTALE). From a practical standpoint, this discovery suggests that print exposure is more influential in the L1 French population compared to L1 Korean speakers. Methodologically, it suggests that the choice of measure is an important consideration when evaluating print exposure in some, but perhaps not all, L2 English language groups. ART, with its more restricted range of potential authors, may not fully reflect the variety of English reading experiences available to French native speakers, whereas the free-form nature of AFT permits them to name a more varied selection of authors. In contrast, native Korean speakers may have more limited exposure to English print and culture generally, as learning English is more often restricted to a classroom setting. As a consequence, in the L1 Korean population the choice of measure is essentially irrelevant, as either AFT or ART serves the same purpose.

We also observed that across both L2 English language groups, AFT was more strongly associated with L2 connectives knowledge than with collocations, and in fact the association between AFT and collocations was only marginal when controlling for LexTALE in the French sample, and non-significant in the Korean sample. There are multiple plausible explanations for this finding. Firstly, collocations are seemingly arbitrary word pairings, therefore they need to be

encountered many times before they are likely to be acquired. Because of this, learning collocations is probably more dependent on the length of time an individual has been learning the target language. Relatedly, the kind of exposure required for learning collocations is likely not restricted to print, as these pairings frequently occur in both casual and formal speech as well. Finally, collocation pairings are virtually limitless in number, making them more difficult for non-native speakers to notice, in addition to being unlikely to be used for targeted instruction. Taken together, it may be that learning collocations requires more lengthy and varied language exposure. Connectives, in contrast, are a more restricted class of words which can be acquired through explicit study. Because connectives perform a particular function by linking separated clauses, their meanings may arguably be best understood through frequent exposure and use—nevertheless, they can still be taught in the classroom. We also consider that certain connectives may be more common (if not entirely restricted) to the written mode, making print exposure more important for their acquisition.

In an eye-tracking study described in Chapter 5, we found that print exposure was associated with predictive looks-to-targets in a visual world paradigm where iconic black-and-white images represented distinct semantic contexts for idiomatic or literal sentences. Among L2 speakers, increased print exposure correlated with increased looks to targets on idiomatic trials, but was correlated with decreased looks to targets on literal trials. Although the former finding aligned with our hypotheses, the latter was an unexpected observation. It might have been reasonable to assume that individuals with greater print exposure would make increased looks to targets regardless of the trial type. Yet it appears that L2

participants with more reading experience sought out potential idiomatic meanings for literal sentences as well, avoiding literal interpretations unless they were certain there was no alternative.

Furthermore, in this eye-tracking study we found the same dissociation between predictions for L1 versus L2 speakers as in the previous studies on connectives and collocations. That is, compared to AFT, ART was more highly correlated with selection accuracy scores on idiomatic trials among L1 English speakers, whereas the reverse was true for L2 speakers. A similar finding was shown for anticipatory looks to targets, with ART scores predicting increased target preference on idiomatic trials in L1 compared to L2 participants, and AFT scores predicting the same in L2 compared to L1 participants. In fact, when controlling for both print exposure measures, the effect of ART on idiomatic trials became significantly negative in L2. However, it is important to point out that when considered alone, ART was still associated with predictive processing of idioms in L2. Instead, the diminished effect of ART in these combined models suggests that ART is uninformative only when accounting for AFT, reflecting the comparatively greater importance for the recall measure of print exposure in L2 participants.

6.2.2 Conclusions

Comparing across studies, we discovered that compared to the Author Recognition Test, the Author Fluency Task is more useful as a proxy measure of print exposure in some L2 English populations, and is equivalent to ART in others. Specifically, AFT was more highly correlated with outcome measures than was ART for L1

French speakers compared to Korean natives, where AFT and ART were equivalent. In the Limitations section below, I describe some considerations for why this might be.

Additionally, it may be that some of the differences between AFT and ART are attributable to the kinds of memory recruited for each measure. Specifically, AFT is an active recall measure which may require more explicit memory, whereas ART is a more passive recognition measure. Although ART author selections are made explicitly (in these studies, using a keyboard to indicate yes or no for author recognition), name recognition does not require an extensive search of memories related to reading. Because of this, we deem that names which are recalled on AFT are more likely to reflect an individual's primary print exposure (i.e. actual reading experience) compared to ART, which may be more indicative of secondary print exposure (i.e. second-hand or cultural exposure to print).

Relatedly, it may also be that the explicit memory requirements of AFT are more relevant for the discourse connectives task. This task requires participants to evaluate each option and integrate it into the sentence frame, evaluating the relationship between clauses and whether the logical relationship holds with the inclusion of each candidate. Nevertheless, controlling for individual differences in verbal fluency using additional semantic fluency tasks (e.g., listing groceries or public figures) did not render insignificant the variance explained by AFT in our models predicting both connectives and collocations scores. From this, we suspect that the constraining factor for AFT is essentially author knowledge, which serves as an appropriate proxy for print exposure.

Finally, as a psychometric measure, AFT shows good convergent validity in both L1 and L2, correlating well with theoretically related measures of language and vocabulary knowledge. Equally important, AFT also shows poor divergent validity, as it does not correlate well with unrelated measures such as task motivation. AFT has also shown acceptable test-retest reliability, although this requires replication with larger sample sizes. Ultimately, we conclude that the AFT is a useful supplement to ART for assessing print exposure, and should be considered as a reasonable alternative in L2 populations.

6.2.3 Limitations and future directions

Although the findings of this thesis suggest that print exposure is associated with L2 vocabulary knowledge, they do not represent a major paradigm shift for understanding second language acquisition. Evidently, input and experience play important roles for L2 learning, and this has been emphasised in the literature for decades. However, the specific contributions of L2 reading experience, as opposed to other forms of language learning, have not always been clear. These findings reinforce the notion that print exposure, which is a significant influence for L1 learning, is also relevant in L2. Moreover, reading may be particularly important for learning more advanced vocabulary, specifically formulaic language.

Similarly, the Author Fluency Task is essentially an iteration of the logic of the Author Recognition Test. With ART, we have known for decades that knowledge of authors is a reliable proxy index for print exposure, and AFT seems at first to simply find a new route to a familiar destination. Yet the two tasks are

clearly distinct, both with respect to the kinds of memory involved and the predictions each measure makes. As mentioned in the previous section, compared to ART, AFT predicted additional variance in connectives and collocations scores for the L1 French sample, but not L1 Korean speakers. One potential explanation is that the fluency task is more discriminative for populations which are more linguistically and culturally proximal to English (e.g. French) compared to those which are more distant (e.g. Korean).

Furthermore, although the validity of ART has been questioned in L2, it was still moderately and reliably associated with formulaic vocabulary in both L1 French and L1 Korean speakers of English as L2. In light of this, researchers should consider the population under study, as well as the specific research question, before selecting which measure to use. Although AFT may explain additional variance beyond ART in some populations, it may not necessarily be the case that AFT is always a superior measure. For example, if the research question involves making cross-linguistic comparisons between multiple L2 groups, ART could be more reliable when used within a single population of L2 speakers, yet may also be less reliable when comparing between L2 groups, since cultural exposure to the target language may vary. However, determining which measure to use and when will require further replication in additional language groups beyond the French and Korean natives tested here.

Relatedly, one of the aims of this thesis was to develop a more reliable measure of print exposure in L2. Despite this goal, it seems unlikely that AFT offers greater precision than ART. In fact, reliability estimates for ART in the L2 populations were very high, whereas estimates for AFT were comparatively

moderate. Nevertheless, the nature of the print exposure measures required different reliability measures. Specifically, ART used KR-20 and ICC2k indices which examine the consistency of response patterns across participants, whereas AFT resorted to test-retest reliability. This makes it difficult to compare the measures directly. Naturally, a participant's responses will likely see significantly more variation when retested on a different day (or even week or month) compared to responses collected during a single session. We were also unable to recruit a large number of participants to be re-tested on AFT, which may have further reduced its reliability. In retrospect, measuring test-retest reliability for ART as well may have enabled more direct comparisons, and this could be a consideration for future work. However, because the list of names on ART is always the same for each participant, and recognition imposes a relatively low memory threshold, reliability will likely be inherently higher for ART. Although this might seem like a point in favour of using ART, it is worth considering that this reliability does not necessarily imply validity in L2 populations.

Future work might also evaluate the patterns of responses on AFT, as they could prove to be relevant to models of lexical access and memory. Briefly touched on in Chapter 1, semantic foraging describes how participants in verbal fluency tasks tend to "cluster" together related items into groups, exhausting these category members before "switching" to a different category which may prove to be more productive. Although the possibility was never formally examined for this thesis, exploratory evidence suggested that some participants may have used this semantic foraging strategy. As an initial first step to evaluate the relative similarity of authors, I coded a feature matrix for each author named on the AFT using available

biographical, demographic, and genre information. For each participant, I then looked at the order in which each author was named. Table 6.2.1 shows example AFT data from an L1 French participant who named a total of 11 authors in the study in Chapter 3.

Table 6.2.1: Example Author Fluency Task data from an L1 French / L2 English participant, arranged by selection order. Information about each author appears to suggest two distinct clusters of related names, consistent with a semantic foraging account of memory.

Order	Author Name	Genre(s)	Gender	Year of Birth	Cluster
1	Stephen King	Horror / Fantasy	Male	1947	1
2	J.R.R. Tolkien	Fantasy	Male	1892	1
3	Frank Herbert	Sci-Fi / Fantasy	Male	1920	1
4	Neil Gaiman	Sci-Fi / Fantasy	Male	1960	1
5	Terry Pratchett	Sci-Fi / Fantasy	Male	1948	1
6	Philip K. Dick	Sci-Fi / Fantasy	Male	1928	1
7	Dan Simmons	Sci-Fi / Fantasy	Male	1948	1
8	Emily Brontë	Literary	Female	1818	2
9	Charlotte Brontë	Literary	Female	1816	2
10	Mary Shelley	Literary / Horror	Female	1797	2
11	George Orwell	Literary / Sci-Fi	Male	1903	2

Prima facie, it appears that there are two distinct clusters in these data. Namely, Cluster 1 is comprised mostly of male sci-fi/fantasy authors writing from the 1940s onwards, whereas Cluster 2 features mostly female authors who were writing literary fiction in the 1800s or early 1900s. Unfortunately, it is difficult to assess the extent to which individuals use semantic foraging strategies because very few

participants produce a sufficient number of authors which would make these analyses possible in the first place. Additionally, it is unclear how to operationalise the degree of similarity between multiple authors. In a test such as ART where items form a restricted range, this is relatively straightforward to accomplish using dimensionality reduction techniques such as factor analysis (e.g., Moore & Gordon, 2015) since they rely on identifying latent response patterns across all items. In an open format task such as AFT, however, the range of potential authors is boundless, and a slightly different approach is required. To examine these questions, we considered the use of techniques which enable quantitative analyses of qualitative data. Some examples of these methods are hierarchical clustering (Henry et al., 2015) and natural language processing techniques such as a vector space approach (e.g., Hollis & Westbury, 2016; Mikolov et al., 2013), although these were not explored for this thesis. However, we briefly explored the use of multiple correspondence analysis (MCA; Le Roux & Rouanet, 2010a, 2010b; Mori et al., 2016). This is a dimensionality reduction technique that analyses sets of categorical variables and computes the distances between them using numerical values which represent the strength of underlying associations. With this technique, it was possible to determine the degree of similarity between authors using our feature matrix. However, it was decided this was not directly relevant for the aims of this thesis, and ultimately, work in this direction was abandoned. Even without these more advanced techniques however, a cursory (though admittedly subjective) exploration of the observations from many of the higher-scoring participants suggests that related author names tend to co-occur relatively frequently, with one name priming another. The operating assumption is that these participants may

be attempting to organise their author knowledge into meaningful categories, creating the impression of a "mental bookshelf". This is similar to findings from semantic fluency studies using grocery items which show that many participants cluster products together as they would be found in particular aisles. However, future studies may wish to explore these data with a more objective approach to quantifying the similarities between authors, in order to make specific, testable predictions.

On a related point, Chapter 2 included exploratory analyses of the relative proportions of fiction and non-fiction, as well as male and female authors provided by male and female participants on AFT. We found that male respondents produced significantly more male and non-fiction author names compared to female counterparts, aligning with media and consumer habits reports which suggested this may be the case. However, this pilot study only recruited 11 male participants, so any conclusions should be considered tentative. One possibility for extending these results on an exploratory basis would be to combine the AFT data from subsequent chapters and run the same statistical tests with a much larger sample size, and to determine if these gender and genre biases vary by language group or by age. It would also be worthwhile to compare these findings to ART, where the gender distribution of authors ideally ought to be more equitable (although the recent ART from Vermeiren et al., 2022 used in this thesis contains 35 male authors and just 25 female authors). For example, if male readers are less familiar with female authors, are they also at a disadvantage when evaluated using a test such as ART? We suspect this is unlikely, given that there is generally no difference in ART scores between genders. Even so, it may still be the case that male

respondents tend to recognise fewer female authors on ART. If there is no difference between male and female response patterns on ART—and if there is on AFT—it would reinforce the notion that ART is more likely to index general cultural exposure rather than individual reading experience.

As mentioned throughout, this thesis is primarily concerned with assessing print exposure using knowledge of formulaic language in L2. However, this was not always the plan. Initially, we trialled using other existing outcome measures such as the synonym selection and sentence completion tasks from Mar and Rain (2015), the spelling recognition task (Andrews & Hersch, 2010), and the vocabulary size test (P. Nation, 2012; P. Nation & Beglar, 2007), among others. These are ingeniously designed and well-validated vocabulary measures, but we eventually chose formulaic language, for the simple reason that this kind of vocabulary is often not taught explicitly, and may be less likely to be tested. Of course, there are many kinds of phrasal units in English, and determining which to use introduced several researcher degrees of freedom. Ultimately, we arrived at using connectives, collocations, and idioms. Connectives were selected because they are a more restricted class of words that can be acquired through explicit learning, although their exact functions may vary somewhat when translated between L1 and L2, and may be best learned through use. Collocations were chosen because they are, in contrast to connectives, virtually infinite, making experience likely even more important for picking up on the statistical regularities of these word pairings. Idioms were selected because they require learners to chunk together units of words to understand the "non-decomposable" figurative meaning. The reasoning was that these vocabulary items are all particularly challenging for L2 speakers, and we

suspected that part of this difficulty stems from a lack of L2 print exposure. Thus, we reasoned that this would make scores on these vocabulary tasks especially useful as a correlate of a potential print exposure measure.

Admittedly, these measures have limitations of their own to consider, both inherent and in terms of implementation. With the collocations task, we did not consider the role of congruency—that is, whether or not a particular L2 word pair is equivalent to one in L1. Evidence shows that congruent collocations are processed faster and more accurately (Szudarski, 2017; Szudarski & Conklin, 2014; Vu & Peters, 2023), and this factor may have explained some of the differences between L1 French and L1 Korean performance on certain items. In the eye-tracking experiment, the sentences used for the idioms task were not validated by other research, since I wrote them from scratch for this study. Again, I maintain that I did all that was feasible during piloting to validate that the sentences were easy to understand for both L1 and L2 speakers. More importantly, the final results also show high selection accuracy on the visual world paradigm. Even so, I suspect there were some sentences which sounded more natural than others, as I may have tried too hard to match the idiomatic and literal versions of the sentences on number of words in each, as well as trying to keep the sentences themselves broadly similar. As with any other measure, of course, developing this visual idiom task will require successive refinements, but it serves as a convincing test case of its general utility.

The eye-tracking study in Chapter 5 measured knowledge of idioms using a visual world paradigm in which images represented distinct semantic contexts. This was a novel approach; to our knowledge, no other study has attempted a similar method. This design proved to be a challenge, as each image needed to be carefully

constructed to ensure it would uniquely specify the context indicated in its sentence. Equally important was to avoid any unintended semantic associations between the sentence and unrelated distractor images. Initially, I had intended to draw each of the images myself to ensure I had some control over these implicit associations. With time, however, I realised that this would have been an enormous undertaking, and one which would have delayed my thesis significantly. Thankfully, I was able to use newly developed image generation software, which spared my brain and wrist further toil. Even so, this software was not without its own challenges. As described in Chapter 5, the large language model would rarely produce exactly what I had in mind, and in most cases, images had to be modified to correct rendering issues or to clarify relationships between characters and objects.

For this study, there was also a risk of some degree of subjectivity in terms of what exactly was conveyed by each image. Yet I would argue that I was uniquely well-situated to design this particular kind of experimental study. In the past, I have had the opportunity to work as an illustrator and a cartoonist, creating storyboards, graphic novels, and animation sequences. Because of this, I have spent many years thinking about what sort of information can be expressed through very simple line drawings. I would like to believe that this experience has made me especially sensitive to the variety of (mis)interpretations which readers might have when "reading" images. Many of the potential interpretations relate to remarkably complex theory of mind knowledge about what a character knows or how they are feeling. Although the drawings themselves may be two-dimensional, readers willingly attribute comic characters with three-dimensional traits. Often, the angle

of a raised eyebrow, or the severity of sine-wave squiggle of a mouth can dramatically alter the inferences readers make about the inner workings of a character's mind. If these variables aren't controlled for, it can render the intended reading a failure, leading to confusion and frustration. For a comic artist, unclear visual information is the equivalent of a garden-path sentence. Although I am certainly no Will Eisner or Jeff Smith, I tried to employ the familiar tricks of the artist's trade to evade these problems, and ran pilot tests to ensure that response accuracy was high for both L1 and L2 speakers.

Another consideration regarding the visual world paradigm relates to the density of information in each image. That is, the images vary with respect to the number of lines or strokes required to produce them. If one image is significantly more visually dense compared to competitors, it may be distracting, which may impel participants to look at it more often (consistent with the visual salience hypothesis, e.g., Gao et al., 2008). On the other hand, students of visual design and advertising are typically taught that the eye and brain are attracted to white (or "negative", or "empty") space (Olsen et al., 2012), and compositions are generally designed to exploit this. Thus, it is important to try to balance the amount of visual information both within and between each image, to the extent that this is possible. In general, I prioritised the principle which is often termed "economy of line", or the parsimonious or minimal use of strokes to convey the required information in a two-dimensional drawing. Still, I recognise that there were likely certain trials where the visual information in each image was not precisely balanced. I suspect it may also be true that sometimes the distractor images were less interesting overall compared to the target and foil images which were the focus of the study. Although

the concept of visual interestingness will likely prove difficult to operationalise, it may be worth considering in a similar or replication experiment.

6.3 Afterword

In this final section, I would like to close with a few broader reflections about the role of print exposure in the development of language, and our knowledge of other minds. I have tried to approach these last pages as a means of clarifying, both scientifically and philosophically, some of my own conclusions and intuitions. Because of its more conversational and personal nature, I have opted to include this section at the end.

6.3.1 Reflections on the use of generative AI tools

I feel compelled to note one aspect of this research which is not solely methodological, but also moral. The image creation software we used for the study in Chapter 5, DALL-E 3 (OpenAI, 2024), is built using "generative AI" large language models. The companies behind these technologies trawl and scrape the

Internet for available data, including text, images, and videos, all of which are used as training data. This information, and the creative output of millions of writers, journalists, programmers, scientists, and artists, is used without implied or explicit consent. Companies such as OpenAI argue that this plundering is necessary to make their technology possible, and that the text and images which their models produce are fundamentally new creations. Yet many argue that these technologies are little more than elaborate plagiarism machines, a brazen act of copyright infringement on a massive scale (including some former employees; Metz, 2024). I'm not certain these detractors are mistaken. Still, I have attempted to rationalise the use of this technology by considering that I am perhaps only sparing myself some time and labour which could be better spent elsewhere as I continue work on my thesis. Ideally of course, I would have paid another artist to draw these images, but this would require time to coordinate and additional funding, both of which are in short supply for post-graduates. In any case, the use of these technologies may give some impression of tacit approval, as though I believe the end justifies the means. This is not my intention, and I recognise that this is something of a morally grey area.

Understandably, many fields are concerned about the use of AI to replace the same creative and technical professionals whose work has been siphoned into these models, and I share this concern. Even so, I want to note that there are doubts whether generative AI models of this kind—which, at the time of writing, still cannot reliably count the number of Rs in the word "strawberry" (Silberling, 2024)—will ever independently produce a great (or even consistently coherent) work of literature, or a profound piece of art. This is not to suggest that they

cannot be useful, or indeed harmful; they will continue to be both. However, I suspect that large language models are not a path towards "artificial intelligence" in any meaningful sense—that is, a computer system that reproduces the mechanisms which give rise to complex thought and consciousness—because their fundamental assumptions about cognition are mistaken. Whether intentional or not on the part of their creators, these models operate as though language is the basis of cognition, rather than the other way around. Yet the chief function of language is to communicate with others, and by doing so, to share some insight into our own minds, and this perspective underscores the social and interactive purpose of language (Fedorenko et al., 2024; M. Johnson, 2018). Some have convincingly argued that there is a kind of "literate thinking" which enables higher-order reasoning and farsightedness (Kolinsky & Morais, 2018; Morais & Kolinsky, 2021a, 2021b), yet literacy is a relatively recent cultural invention; a new addition to our linguistic toolkit. Nevertheless, I suspect that due to the perceived higher status of writing, and its utility for clarifying our own ideas, we have become accustomed to thinking about language in a disembodied sense, as though it exists apart from any individual who produced it, or any receiver to interpret it. Yet language and cognition both are rooted in embodied experiences and their associated qualia (M. Johnson, 2018; Lakoff, 2012), which these models do not have. A large language model has no internal observer, and no subjective experience; there is no one on the other end with whom we are speaking. Even so, we might charitably refer to what these models do as "thinking", at least in the behavioural or performance sense as Turing (1950) originally conceived it, and certainly they are performing complex internal operations which aren't always obvious to us. But these models have no

interiority or intentionality, and they cannot consider any personal lived experience. Instead, they are pattern detectors, asking only the probability of *B* given *A*. In other words, these models may possess virtually infinite knowledge, but they have no understanding. Without understanding, how can there be creative insight, or emotional resonance? Consequently, this is likely something which will not be overcome just by feeding the same models more data, as though we might simply turn up the valve on the "word hose".[25] Granted, many will rightly point out that much of the creative output of humans is similarly iterative (if not derivative)—the subtext being that perhaps little of value will be lost in the automation of such work. Yet art and stories are not solely about entertainment value or diversion; they are fundamental for understanding ourselves and others. This might seem like a lofty, humanitarian aim, but because empathy and theory of mind are important aspects of language development, it is arguably practical as well.

Of course, generative AI systems also present new practical opportunities for second language learning as well. Manifested as chatbots, these large language models can be "prompt-engineered" to function as personal language tutors—ones which are keen to discuss whatever topic the user may choose, and which have infinite patience for correcting grammatical errors. With newer models, users can even interact naturally using their voices, providing the opportunity to develop listening skills as well. This may prove to be a groundbreaking technology, especially in the developing world where language lessons may be prohibitively

[25] Then again, if you had asked me ten years ago if today we would have chatbots which could create pictures and video, discuss poetry, and pass the LSAT examinations, I would have probably guessed not.

expensive. However, there are some potential points of failure which may limit their usefulness. For example, the models must first convert speech into text before parsing, but accented speech is more likely to be misinterpreted by automatic speech recognition systems (Wassink et al., 2022), potentially providing inaccurate transcriptions to the model before the user is able to correct it. Clearly, this limits the technology's usefulness for absolute beginners. Misunderstandings are typical in L1/L2 interactions as well, but speakers can evaluate their confidence in what they have heard and request clarification, or for their interlocutors to repeat as needed. In contrast, LLMs will tend to cobble together a specious response to virtually any input without evaluating for reasonableness (Helal et al., 2024). This may lead to instances where the model responds with irrelevant information, leading to more confusion. Again, similar mix-ups are not uncommon in L1/L2 interactions between humans, but with shared context and a little patience, they are usually surmountable.

Additionally, chatbots can be prone to "hallucinations", or instances where they invent certain facts, which they share with undue confidence (Athaluri et al., 2023; Emsley, 2023; Goddard, 2023). Many popular languages are well-documented online, meaning this is unlikely to pose a problem specifically for learning accurate information about these languages. Less common languages, however, having smaller or even non-existent web presences, will naturally be less reliable, but it is unlikely that the chatbot will be willing to recognise and acknowledge the gaps in its own knowledge as we might expect a human tutor to do. A related problem surrounds the tendency of LLMs to become "confused", or more accurately, to provide increasingly inappropriate responses as a dialogue with a user continues

(Associated Press, 2023; Roose, 2023). Because of this, language learning services which are designed to use this technology typically feature guardrails to prevent users from venturing too far astray from a particular topic. These include measures such as preventing conversations from exceeding a given length, as well as restricting certain topics which are likely to yield insensitive or impolite replies. Without question, these are all sensible and responsible protections to implement. Anecdotally however, after trying out many of these services personally, I came away with the impression that the sorts of conversations which are possible in these frameworks are so rigidly regimented that they often feel less like communicating naturally and spontaneously with another speaker, and more like a traditional grammar exercise—but one which is far less focused. Moreover, to "improve their service" (and to ensure that their services aren't being exploited) most AI learning sites advise users that conversations may be reviewed, warning that no private information should be provided. Again, this is a reasonable, even boilerplate proviso to include in a site's terms and conditions, and frankly, it is good advice for using the Internet generally. Yet such messages give the unsettling impression that the user is in effect under a microscope, and that everything being said can (and will) be used as more grist for the AI mill. Unfortunately, I suspect that for most people, this knowledge will be stifling, precluding a relaxed and informal environment which is critical for language learning. Still, I am hopeful that we will find a way of making this technology more broadly accessible and, perhaps with appropriate legislation, more ethical as well.

6.3.2 Language and other minds

As has been discussed at length, this thesis has primarily explored which factors influence second language learning. Although the experimental work is correlational, it forms a consistent pattern of association between print exposure and formulaic vocabulary in L2. Readers who know more L2 authors demonstrate greater knowledge about words and phrases, reflecting the "lexical legacy" of their diverse reading experiences (K. Nation, 2017). From a theoretical perspective, the results are consistent with emergentist or usage-based theories which emphasise the importance of experience for language acquisition. The findings may also suggest that the way children and adults acquire language is not fundamentally different. This may feel counterintuitive—after all, whereas children are typically regarded as preternaturally gifted language learners, most of us find language learning in adulthood to be frustrating and unproductive. But I suspect the most important advantage children have over adults is not that they possess some kind of linguistic superpower, but rather that they are rarely self-conscious about language, and they are willing to play with it. Children will happily pick up a regular past tense morpheme and jam it onto an irregular verb, like toy wheels stuck to a LEGO house—unbothered that they have never heard anyone say "runned" or "fighted" before. They will also happily pair off words which have never been formally introduced, like two strangers sat next to each other at a dinner party. And they have no compunction whatsoever about using a word or phrase before they have looked up its definition in the dictionary, because this is not where words live. Contrast this freewheeling attitude with how adolescents and adults generally approach language learning. In the book "Hokkaido Highway Blues" (2003), the

Canadian writer Will Ferguson provides a firsthand account of teaching English overseas, remarking:

> "My students in Japan were determined to reduce English to mathematical dictums that could then be reassembled. One student, who was a diligent pupil but refused to speak English with me in class, said with perfect sincerity, "It's just that I hate to make mistakes. So, first I will become fluent in English and then I will speak it." When I tried to explain to him that learning a language was a process and that making mistakes was a necessary, even desirable aspect of it, he politely dismissed my suggestions as being eccentric." (pp. 286-287).

If language is a kind of game, then we ought to consider how we learn to play games, both as children and as adults. Most of us are familiar with the experience of having the rules of a new board game tediously explained to us. After a well-meaning friend has droned interminably about the various pieces, turn progressions, and conditions for victory, we ultimately have to admit we didn't understand any of what we've just heard. We throw up our hands and say, "let's just play a round, and I'll see how it works". Of course, some part of the explanation of the rules has likely taken root, priming us for our performance in the first round, yet only by playing the game does it become knowledge. In the Sufi faith, this idea is encapsulated in the saying, "you can only learn what you have already learned" (Kaufmann, 2017). In the same way, language knowledge must be consolidated through use—through play. Yet so much second language learning, at least traditionally, has been approached as an intellectual exercise, like learning a board game by first memorising the rules before ever playing the game. Because of this

focus on the supposedly rational "rules" of language, native speakers sometimes mistakenly assume that a second language speaker who makes grammatical errors is lacking intelligence. Hopefully it is obvious by now that this is utterly false, but to illustrate, I will have to invoke the name of one of my favourite authors.

Among the most celebrated English writers, in another life, Virginia Woolf might also have made a preeminent cognitive scientist. In a radio broadcast for the BBC in 1937 (later published posthumously as an essay), she mused about observing words in their natural habitat, vividly conjuring the image of looking down into "that deep, dark and only fitfully illuminated cavern in which they live — the mind" (Woolf, 1937, 1942, 2012). Here, she recalls the early psychologist William James, who compared introspective analysis in psychology to "trying to turn up the gas quickly enough to see how the darkness looks" (W. James, 1884; Mendelsund, 2014). Woolf reflects that when we try to translate our thoughts into language, words themselves (imbued here with anthropomorphic volition) are not concerned about which particular lexical items are plucked up from that deep, dark cavern—as though only one or another is fit to purpose—only that we "think and feel" before using them, in an effort to faithfully express our intent. Ostensibly an essay about the work of a writer, Woolf aptly distinguishes between cognition and language, noting that we often feel things for which we may never find the words.

Like the rest of us, second-language speakers may also occasionally struggle to find the "right words". Yet the language of non-native speakers, even those who are beginners, is not somehow "deficient", nor is their grammar "broken"—on the contrary, it is working exactly as intended. Knowledge of a second language is informed by the same factors which influence mastery over our mother

tongues, including experience, context, salience, frequency and diversity, background knowledge, and socio-cognitive factors such as empathy and theory of mind. Reading, which engages and informs all of these, is an especially important source of vocabulary, both in L1 and L2. Granted, L2 print exposure essentially predicts how closely second language speakers' internal models of language resemble those of L1 peers. This is not to suggest, however, that "native-like speech" is the desired outcome for all non-native speakers. Even leaving aside the dubious and moralising tone of such an idea, it is unclear how we might define native speech. Even amongst L1 speakers, evidence shows that there is considerable variation in performance on proficiency measures and knowledge of certain grammatical structures (Dąbrowska, 2015), demonstrating that there is no single, monolithic grammar which all native speakers share. We can safely infer, then, that the "rules" of grammar—rather than being handed down from on high on stone tablets—are little more than the predilections of prestige varieties of language, themselves the result of arbitrary happenstance, an accumulation of convention, style, tone, and register. Rather than one solitary English, there are many Englishes, each occupying the minds of its speakers, bustling with those promiscuous and roving inhabitants, those words which intermingle and intermarry freely between languages and dialects without shame or prejudice. The etymology and orthography of English words bear witness to these messy histories, leaving echoes and palimpsests of languages which came before our own "magnificent bastard tongue" (McWhorter, 2008).

With all this in mind, it is worth reiterating that despite the present focus on formulaic language, there is nothing superior—morally or otherwise—about

expressing a particular thought using one set of words or another. Neither can there be anything pure or impure about language. Non-native speakers may occasionally express their thoughts using atypical formulations, and naturally, this may place additional processing demands on the receiving end of the message. Rather than being a hindrance however, this variation is a necessary ingredient for a living language—after all, the only languages in which rules and conventions are firmly fixed are those for which rigor mortis has already set in. Words, Woolf writes,

> "[...] mean one thing to one person, another thing to another person; they are unintelligible to one generation, plain as a pikestaff to the next. And it is because of this complexity that they survive. Perhaps then one reason why we have no great poet, novelist or critic writing to-day is that we refuse words their liberty.[26] We pin them down to one meaning, their useful meaning, the meaning which makes us catch the train, the meaning which makes us pass the examination. And when words are pinned down they fold their wings and die."

Literary fiction, including Woolf's own writing of course, is renowned for its creative use of language. Compared to second language speakers however, when an author's choice of language strikes us as unusual, they are more often given the benefit of the doubt. Great writing tends to avoid the formulaic and cliché, and this novelty causes us to take notice of the exceptions which prove the rules. How similar this is to the speech of second language learners—native speakers rarely interrogate why we say something one way until we hear it said another. This noticing is an

[26] I hope the reader will appreciate the humour in Woolf believing there was no great writer alive in her time.

important ingredient for learning (Andringa, 2020; Bishop, 2004; Schmidt, 1990)—and when our expectations about language are subverted, like an unexpected plot twist, readers and speakers must slow down to reinterpret or reconsider what we've just read or heard, forcing us to sip and savour rather than gorge and gobble.

Yet this linguistic ingenuity is arguably not the principal defining feature of literature (neither is "books they make you read in secondary school", though it isn't a bad starting point). Literary fiction is often positioned in contrast to "genre fiction" such as sci-fi, fantasy, mystery, or romance, which are all equally legitimate forms of reading, and important sources of language knowledge in their own right. It has been suggested however, that what is unique about literary fiction is that it allows us to not simply encounter other minds, but to inhabit them, quite literally, by experiencing a kind of "literary possession" (Storr, 2021). Granted, genre fiction also encourages readers to consider the perspectives and motivations of different characters. Yet these categories have traditionally focused less on the interiority of how events are perceived and processed by the narrator(s) so much as the importance of external events which drive the plot forward. Thus, the Mountie always gets his man, and so too Bridget Jones (we hope).

In contrast, how should one summarise Woolf's modernist fever dream of a novel *To the Lighthouse*? A strict synopsis of plot points might give the impression that it is at bottom a story about a dysfunctional family failing to plan a holiday. Granted, this would mean ignoring the complex character dynamics which unravel through stream-of-consciousness narration, flitting briskly from one perspective to another with no warning. This characteristic gives the novel its richness and layered insight, yet it would be difficult, if not impossible, to strip

mine the contents of these passages and filter them into something worthy of a book flap. Indeed, a common refrain (or rather, complaint) about literary fiction is that "nothing happens". More often, what is meant by this is that the internal conflict which is central to the story fails to materialise into plot elements, or external action. In this way, literature is something like our internal thoughts which are rarely summoned into words—the real action is going on inside.

Admittedly, it can be extremely challenging to endure a story when it isn't clear what the stakes are, or when we can't predict the likely outcomes. Reading through an arduous work of fiction requires a high "tolerance of ambiguity"; that stubborn willingness to persist even when things aren't perfectly clear. Similarly, second language learning requires significant motivation to persevere in the face of much uncertainty, and evidence shows this trait is also implicated in multilingualism (Dewaele & Wei, 2013; Purpuri et al., 2023). Just as we try to infer the motivations of characters in a book, successfully conversing in a second language is often a question of inferring the internal workings of other minds, and these can seem opaque or impenetrable. Clearly, the "useful meanings" of words—those meanings which help us "to catch the train, or to pass the examination"—are a necessary ingredient for language competence. Yet proficiency requires more than this. It is equally important to become familiar with the places where words live; the minds of other speakers. Because our words bear the lexical legacy of our accumulated exposure to language, we see a flash of torchlight on these dark caverns with every spoken exchange or written paragraph. How could it be any other way? To paraphrase an Arabic proverb, "a glass can only spill what it contains".

Bibliography

Achard, M., & Niemeier, S. (2004). *Cognitive linguistics, second language acquisition, and foreign language teaching*. Mouton de Gruyter.

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Amenta, S., Hasenäcker, J., Crepaldi, D., & Marelli, M. (2023). Prediction at the intersection of sentence context and word form: Evidence from eye-

movements and self-paced reading. *Psychonomic Bulletin & Review*, *30*(3), 1081–1092. https://doi.org/10.3758/s13423-022-02223-9

Anderson, D. R., & Burnham, K. P. (2002). Avoiding Pitfalls When Using Information-Theoretic Methods. *The Journal of Wildlife Management*, *66*(3), 912–918. https://doi.org/10.2307/3803155

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719. https://doi.org/10.1037/0033-295X.96.4.703

Andersson, M., & Sundberg, R. (2021). Subjectivity (Re)visited: A Corpus Study of English Forward Causal Connectives in Different Domains of Spoken and Written Language. *Discourse Processes*, *58*(3), 260–292. https://doi.org/10.1080/0163853X.2020.1847581

Andersson, R., Nyström, M., & Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research*, *3*(3). https://doi.org/10.16910/jemr.3.3.6

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, *139*(2), 299–318. https://doi.org/10.1037/a0018366

Andringa, S. (2020). The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research*, *36*(3), 335–357. https://doi.org/10.1177/0267658320915502

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder.

Behavior Research Methods, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arnon, I., & Christiansen, M. H. (2017). The Role of Multiword Building Blocks in Explaining L1–L2 Differences. *Topics in Cognitive Science*, *9*(3), 621–636. https://doi.org/10.1111/tops.12271

Associated Press. (2023, February 22). Microsoft curbs Bing's bad behaviour as it debuts AI chatbot on phones. *CBC News*. https://www.cbc.ca/news/business/microsoft-bing-chatbot-mobile-phones-1.6756110

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, *15*(4), e37432. https://doi.org/10.7759/cureus.37432

Auxier, B., Stewart, D., Bucaille, A., & Westcott, K. (2021, December 1). *The gender gap in reading: Boy meets book, boy loses book, boy never gets book back*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/gender-gap-in-reading.html

Baddeley, A. D. (1972). Retrieval rules and semantic coding in short-term memory. *Psychological Bulletin*, *78*(5), 379–385. https://doi.org/10.1037/h0033477

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Elsevier. https://doi.org/10.1016/S0079-7421(08)60452-1

Baker, D. L., Stoolmiller, M., Good Iii, R. H., & Baker, S. K. (2011). Effect of Reading Comprehension on Passage Fluency in Spanish and English for Second-Grade English Learners. *School Psychology Review*, *40*(3), 331–351. https://doi.org/10.1080/02796015.2011.12087702

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Baron, N. S. (2016, July 20). Why Digital Reading Is No Substitute for Print. *The New Republic*. https://newrepublic.com/article/135326/digital-reading-no-substitute-print

Baron, N. S. (2020). Digital Reading: A Research Assessment. In *Handbook of Reading Research, Volume V*. Routledge.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. In *International Review of Research in Mental Retardation* (Vol. 23, pp. 169–184). Elsevier. https://doi.org/10.1016/S0074-7750(00)80010-5

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J., Gann, T. M., & Pierce, R. S. (2011). Anticipatory baseline effects and information integration in visual world studies. *Acta Psychologica*, *137*(2), 201–207. https://doi.org/10.1016/j.actpsy.2010.09.011

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2022). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4* (Version 1.1-29) [Computer software]. https://CRAN.R-project.org/package=lme4

Beck, S. D., & Weber, A. (2016). Bilingual and Monolingual Idiom Processing Is Cut from the Same Cloth: The Role of the L1 in Literal and Figurative Meaning Activation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01350

Berends, S. M., Brouwer, S. M., & Sprenger, S. A. (2016). Eye-Tracking and the Visual World Paradigm. In M. S. Schmid, S. M. Berends, C. Bergmann, S. M. Brouwer, N. Meulman, B. J. Seton, S. A. Sprenger, & L. A. Stowe, *Designing Research on Bilingual Development* (pp. 55–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-11529-0_5

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied*, *5*(4), 438–447. https://doi.org/10.1037/1076-898X.5.4.438

Berman, R. A., & Nir, B. (2010). The lexicon in writing–speech-differentiation. *Written Language & Literacy*, *13*(2), 183–205. https://doi.org/10.1075/wll.13.2.01ber

Berthele, R., & Vanhove, J. (2020). What would disprove interdependence? Lessons learned from a study on biliteracy in Portuguese heritage language speakers in Switzerland. *International Journal of Bilingual Education and Bilingualism*, *23*(5), 550–566. https://doi.org/10.1080/13670050.2017.1385590

Biber, D. (1988). *Variation across Speech and Writing* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing Company. https://doi.org/10.1075/scl.23

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of spoken and written English*. John Benjamins Publishing Company.

Bishop, H. (2004). Noticing Formulaic Sequences—A Problem of Measuring the Subjective. *LSO Working Papers in Linguistics*, *4*, 15–19.

Black, J. E. (2019). An IRT Analysis of the Reading the Mind in the Eyes Test. *Journal of Personality Assessment*, *101*(4), 425–433. https://doi.org/10.1080/00223891.2018.1447946

Black, J. E., & Barnes, J. L. (2015). The effects of reading material on social and non-social cognition. *Poetics*, *52*, 32–43. https://doi.org/10.1016/j.poetic.2015.07.001

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. M. Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (1st ed., pp. 41–68). Cambridge University Press. https://doi.org/10.1017/CBO9781139524544.005

Bley-Vroman, R. (1996). What we have to explain in foreign language learning. *Behavioral and Brain Sciences*, *19*(4), 718–718. https://doi.org/10.1017/S0140525X00043570

Bloom, P. (2000). *How Children Learn the Meanings of Words*. The MIT Press. https://doi.org/10.7551/mitpress/3577.001.0001

Bloom, P. (2001). Précis of *How Children Learn the Meanings of Words*. *Behavioral and Brain Sciences*, *24*(6), 1095–1103. https://doi.org/10.1017/S0140525X01000139

Bloom, P. (2005). Word Learning, Intentions, and Discourse. *Journal of the Learning Sciences*, *14*(2), 311–314. https://doi.org/10.1207/s15327809jls1402_8

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Greenwood Press.

Brysbaert, M. (2013). Lextale_FR A Fast, Free, and Efficient Test to Measure Language Proficiency in French. *Psychologica Belgica*, *53*(1), 23. https://doi.org/10.5334/pb-53-1-23

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Sui, L., Dirix, N., & Hintz, F. (2020). Dutch Author Recognition Test. *Journal of Cognition*, *3*(1), 6. https://doi.org/10.5334/joc.95

Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, *49*(2), 772–783. https://doi.org/10.3758/s13428-016-0747-8

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Burt, J. S., & Fury, M. B. (2000). Spelling in adults: The role of reading skills and experience. *Reading and Writing*, *13*(1), 1–30. https://doi.org/10.1023/A:1008071802996

Bybee, J. (2007). Sequentiality as the Basis of Constituent Structure. In J. Bybee, *Frequency of Use and the Organization of Language* (1st ed., pp. 313–335). Oxford University Press; New York. https://doi.org/10.1093/acprof:oso/9780195301571.003.0015

Bybee, J. (2010). *Language, Usage and Cognition* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Caldwell-Harris, C. L. (2021). Frequency effects in reading are powerful – But is contextual diversity the more important variable? *Language and Linguistics Compass*, *15*(12), e12444. https://doi.org/10.1111/lnc3.12444

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition*, *43*(2), 193–205. https://doi.org/10.3758/s13421-014-0462-6

Carey, S. (2010). Beyond Fast Mapping. *Language Learning and Development*, *6*(3), 184–205. https://doi.org/10.1080/15475441.2010.484379

Castano, E., Paladino, M. P., Cadwell, O. G., Cuccio, V., & Perconti, P. (2021). Exposure to Literary Fiction Is Associated With Lower Psychological Essentialism. *Frontiers in Psychology*, *12*, 662940. https://doi.org/10.3389/fpsyg.2021.662940

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cerquiglini, B. (2024). *'La langue anglaise n'existe pas': C'est du français mal prononcé*. Gallimard.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *pwr: Basic Functions for Power Analysis* (Version 1.3-0) [Computer software]. https://CRAN.R-project.org/package=pwr

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651. https://doi.org/10.1207/s15516709cog2605_3

Chang, F., Kidd, E., & Rowland, C. F. (2013). Prediction in processing is a by-product of language learning. *Behavioral and Brain Sciences*, *36*(4), 350–351. https://doi.org/10.1017/S0140525X12002518

Chapple, M., Williams, S., Billington, J., Davis, P., & Corcoran, R. (2021). An analysis of the reading habits of autistic adults compared to neurotypical adults and implications for future interventions. *Research in Developmental Disabilities*, *115*, 104003. https://doi.org/10.1016/j.ridd.2021.104003

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*(1), 143–153. https://doi.org/10.3758/BF03211582

Chen, H. (2022). *VennDiagram: Generate High-Resolution Venn and Euler Plots* (Version 1.7.3) [Computer software]. https://cran.r-project.org/web/packages/VennDiagram/index.html

Chen, S., & Fang, S. (2015). Developing a Chinese version of an Author Recognition Test for college students in Taiwan. *Journal of Research in Reading*, *38*(4), 344–360. https://doi.org/10.1111/1467-9817.12018

Chernova, D. A., & Bakhturina, P. V. (2023). Method of print exposure assessment: Application in psycholinguistics and adaptation for the Russian language.

Vestnik of Saint Petersburg University. Language and Literature, *20*(4), 872–887. https://doi.org/10.21638/spbu09.2023.412

Cheung, A. (2022). *Usage-Inspired Insights into Second Language Learning: A Comparative Review of Usage-Based Studies on Vocabulary Development*. https://doi.org/10.17863/CAM.90578

Chiswick, B. R., & Miller, P. W. (2005). Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages. *Journal of Multilingual and Multicultural Development*, *26*(1), 1–11. https://doi.org/10.1080/14790710508668395

Cho, B.-E. (2004). Issues Concerning Korean Learners of English: English Education in Korea and Some Common Difficulties of Korean Students. *The East Asian Learner*, *1*(2).

Choi, L. J. (2021). 'English is always proportional to one's wealth': English, English language education, and social reproduction in South Korea. *Multilingua*, *40*(1), 87–106. https://doi.org/10.1515/multi-2019-0031

Chomsky, N. (2015). *Syntactic structures* (Repr. der Ausg.'s-Gravenhage, Mouton, 1957). Martino Publ.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Christiansen, M. H., & Chater, N. (2022). *The language game: How improvisation created language and changed the world*. Basic Books.

Church, K. W., & Hanks, P. (2008). Word Association Norms, Mutual Information, and Lexicography. In T. Fontenelle (Ed.), *Practical Lexicography* (pp. 285–

295). Oxford University PressOxford. https://doi.org/10.1093/oso/9780199292332.003.0019

Cieślicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, *22*(2), 115–144. https://doi.org/10.1191/0267658306sr263oa

Cipielewski, J., & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, *54*(1), 74–89. https://doi.org/10.1016/0022-0965(92)90018-2

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford university press.

Clark, E. V. (2020). Conversational Repair and the Acquisition of Language. *Discourse Processes*, *57*(5–6), 441–459. https://doi.org/10.1080/0163853X.2020.1719795

Conklin, K., & Schmitt, N. (2008). Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, *29*(1), 72–89. https://doi.org/10.1093/applin/amm022

Conklin, K., & Schmitt, N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Constantino, R., Lee, S.-Y., Cho, K.-S., & Krashen, S. (1997). Free Voluntary Reading as a Predictor of TOEFL Scores. *Applied Language Learning*, *8*(1), 111–118.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. *Review of Educational Research*, *49*(2), 222. https://doi.org/10.2307/1169960

Cunningham, A. E., & Stanovich, K. E. (1990). Assessing print exposure and orthographic processing skill in children: A quick measure of reading experience. *Journal of Educational Psychology*, *82*(4), 733–740. https://doi.org/10.1037/0022-0663.82.4.733

Cunningham, A. E., & Stanovich, K. E. (1998). What Reading Does for the Mind. *American Educator*, *22*(1–2), 8–15.

Dąbrowska, E. (2014). Words that go together: Measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon*, *9*(3), 401–418. https://doi.org/10.1075/ml.9.3.02dab

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00852

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition*, *178*, 222–235. https://doi.org/10.1016/j.cognition.2018.05.018

Dąbrowska, E., & Divjak, D. (Eds.). (2019). *Cognitive linguistics*. De Gruyter Mouton.

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English

speakers. *Language Sciences*, *28*(6), 604–615. https://doi.org/10.1016/j.langsci.2005.11.014

Davidson, M. M., & Ellis Weismer, S. (2018). A preliminary investigation of parent-reported fiction versus non-fiction book preferences of school-age children with autism spectrum disorder. *Autism & Developmental Language Impairments*, *3*, 2396941518806109. https://doi.org/10.1177/2396941518806109

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*, *1*(1), 9–53. https://doi.org/10.34842/5WE1-YK94

DeVellis, R. F. (2017). *Scale development: Theory and applications* (Fourth edition). SAGE.

Dewaele, J.-M., & Wei, L. (2012). Multilingualism, empathy and multicompetence. *International Journal of Multilingualism*, *9*(4), 352–366. https://doi.org/10.1080/14790718.2012.714380

Dewaele, J.-M., & Wei, L. (2013). Is multilingualism linked to a higher tolerance of ambiguity? *Bilingualism: Language and Cognition*, *16*(1), 231–240. https://doi.org/10.1017/S1366728912000570

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2019). Prediction and integration of semantics during L2 and L1 listening. *Language, Cognition and Neuroscience*, *34*(7), 881–900. https://doi.org/10.1080/23273798.2019.1591469

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2). https://doi.org/10.1515/cllt.2010.006

Dussias, P. E. (2010). Uses of Eye-Tracking Data in Second Language Sentence Processing Research. *Annual Review of Applied Linguistics*, *30*, 149–166. https://doi.org/10.1017/S026719051000005X

EF Education First. (2023). *EF English Proficiency Index*. https://www.ef.co.uk/epi/

Eigsti, I.-M., De Marchena, A. B., Schuh, J. M., & Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders*, *5*(2), 681–691. https://doi.org/10.1016/j.rasd.2010.09.001

Ellis, N. C. (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2012). Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/S0267190512000025

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Ellis, R., Loewen, S., Elder, C., Reinders, H., Erlam, R., & Philp, J. (Eds.). (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual Matters.

Emsley, R. (2023). ChatGPT: These are not hallucinations – they're fabrications and falsifications. *Schizophrenia*, *9*(1), 1–2. https://doi.org/10.1038/s41537-023-00379-4

ESV Online. (2001). *English Standard Version Bible*. https://esv.literalword.com/

ETS. (2024). *TOEFL iBT® Test and Score Data Summary 2023*. https://www.ets.org/pdfs/toefl/toefl-ibt-test-score-data-summary-2023.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in spoken language processing. *Journal of Experimental Psychology: General*, *150*(10), 2167–2174. https://doi.org/10.1037/xge0001042

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*(8017), 575–586. https://doi.org/10.1038/s41586-024-07522-w

Ferguson, W. (2003). *Hokkaido highway blues: Hitchhiking Japan* (Abridged ed). Canongate.

Finney, S. J., Mathers, C. E., & Myers, A. J. (2016). Investigating the Dimensionality of Examinee Motivation across Instruction Conditions in

Low-Stakes Testing Contexts. *Research & Practice in Assessment*, *11*, 5–17. https://doi.org/10.1080/15305058.2015.1034866

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-55. In *Studies in Linguistic Analysis* (pp. 1–31). Basil Blackwell.

Flege, J. E. (2008). Give Input a Chance! In T. Piske & M. Young-Scholten (Eds.), *Input Matters in SLA* (pp. 175–190). Multilingual Matters. https://doi.org/10.21832/9781847691118-012

Flege, J. E. (2019). A non-critical period for second-language learning. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen, *A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn* (pp. 501–541). Aahus University Library. https://doi.org/10.7146/aul.322.218

Fleva, E., Tsimpli, I. M., Fotiadou, G., & Katsiperi, M. (2017). The effect of print exposure upon performance on the Raven Progressive Matrices Test. *Selected Papers on Theoretical and Applied Linguistics*, *22*(0), 133–145. https://doi.org/10.26262/istal.v22i0.5976

Fong, K., Mullin, J. B., & Mar, R. A. (2013). What you read matters: The role of fiction genre in predicting interpersonal sensitivity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 370–376. https://doi.org/10.1037/a0034084

Foster, P. (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*. Routledge.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D.,

Nilsson, H., ... R-Core. (2022). *car: Companion to Applied Regression* (Version 3.0-13) [Computer software]. https://CRAN.R-project.org/package=car

Frick, R. W. (1992). Interestingness. *British Journal of Psychology*, *83*(1), 113–128. https://doi.org/10.1111/j.2044-8295.1992.tb02427.x

Friesen, D. C., Luo, L., Luk, G., & Bialystok, E. (2015). Proficiency and control in verbal fluency performance across the lifespan for monolinguals and bilinguals. *Language, Cognition and Neuroscience*, *30*(3), 238–250. https://doi.org/10.1080/23273798.2014.918630

Friez, C. (2022). *B-Roll | EditMentor Help Center*. https://help.editmentor.com/en/articles/5783539-b-roll

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, *8*(7), 13. https://doi.org/10.1167/8.7.13

Goddard, J. (2023). Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. *The American Journal of Medicine*, *136*(11), 1059–1060. https://doi.org/10.1016/j.amjmed.2023.06.012

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, *16*(4), 562. https://doi.org/10.1037/0894-4105.16.4.562

Goodreads. (2022). https://www.goodreads.com/author/

Gorilla Experiment Builder. (n.d.). *Eye Tracking in Gorilla*. Gorilla Support. Retrieved 25 February 2024, from https://support.gorilla.sc/support/tools/legacy-tools/task-builder-1/eye-tracking#metrics

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced Outlines, Familiarity, and Text Genre on Retention of Prose. *The Journal of Experimental Education*, *48*(4), 281–290. https://doi.org/10.1080/00220973.1980.11011745

Granger, S. (1998). Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford University Press; Oxford. https://doi.org/10.1093/oso/9780198294252.003.007

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439. https://doi.org/10.1038/nature02029

Grüter, T., Rohde, H., & Schafer, A. (2014). The role of discourse-level expectations in non-native speakers' referential choices. *Proceedings of the Annual Boston University Conference on Language Development*. https://par.nsf.gov/biblio/10028988-role-discourse-level-expectations-non-native-speakers-referential-choices

Guiora, A. Z., Brannon, R. C. L., & Dull, C. Y. (1972). Empathy and second language learning. *Language Learning*, *22*(1), 111–130. https://doi.org/10.1111/j.1467-1770.1972.tb00077.x

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, *23*(2), 283–294. https://doi.org/10.1017/S1366728919000026

Hallin, A. E., & Van Lancker Sidtis, D. (2017). A Closer Look at Formulaic Language: Prosodic Characteristics of Swedish Proverbs. *Applied Linguistics*, *38*(1), 68–89. https://doi.org/10.1093/applin/amu078

Hardie, A. (2014). Log Ratio – an informal introduction | ESRC Centre for Corpus Approaches to Social Science (CASS). *ESRC Centre for Corpus Approaches to Social Science (CASS)*. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *The British Journal of Clinical Psychology*, *39*(2), 181–191. https://doi.org/10.1348/014466500163202

Hart, S. A., Little, C., & van Bergen, E. (2021). Nurture might be nature: Cautionary tales and proposed solutions. *Npj Science of Learning*, *6*(1), 1–12. https://doi.org/10.1038/s41539-020-00079-z

Hartig, F. (2022). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models* (Version 0.4.7) [Computer software]. https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., De Geus, E. J. C., Van Beijsterveldt, C. E. M., Bartels, M., Posthuma, D., Boomsma, D. I., Davis, O. S. P., Kovas, Y., Corley, R. P., DeFries, J. C., Hewitt, J. K., Olson, R. K., Rhea, S.-A., Wadsworth, S. J., Iacono, W. G., McGue, M., ... Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry*, *15*(11), 1112–1120. https://doi.org/10.1038/mp.2009.55

Helal, M., Holthaus, P., Lakatos, G., & Amirabdollahian, F. (2024). *Chat Failures and Troubles: Reasons and Solutions* (No. arXiv:2309.03708). arXiv. https://doi.org/10.48550/arXiv.2309.03708

Henry, D., Dymnicki, A. B., Mohatt, N., Allen, J., & Kelly, J. G. (2015). Clustering Methods with Qualitative Data: A Mixed-Methods Approach for Prevention Research with Small Samples. *Prevention Science*, *16*(7), 1007–1016. https://doi.org/10.1007/s11121-015-0561-z

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Hsiao, Y., Banerji, N., & Nation, K. (2021). Boys Write About Boys: Androcentrism in Children's Reading Experience and Its Emergence in Children's Own Writing. *Child Development*, *92*(6), 2194–2204. https://doi.org/10.1111/cdev.13623

Hsiao, Y., Dawson, N. J., Banerji, N., & Nation, K. (2022). The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language*, *50*(3), 1–26. https://doi.org/10.1017/S0305000921000957

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) Learners: Behavioural predictors and neural substrates. *Brain and Language*, *127*(3), 366–376. https://doi.org/10.1016/j.bandl.2012.11.006

Huang, Q. (Helen), & Bolt, D. M. (2023). Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02275-2

Huettig, F., & Pickering, M. J. (2019). Literacy Advantages Beyond Reading: Prediction of Spoken Language. *Trends in Cognitive Sciences*, *23*(6), 464–475. https://doi.org/10.1016/j.tics.2019.03.008

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Hug, M., Jarosch, J., Eichenauer, C., Pennella, S., Kretzschmar, F., & Nicklas, P. (2024). Some students are more equal: Performance in Author Recognition

Test and Title Recognition Test modulated by print exposure and academic background. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02330-y

Ibbotson, P. (2013). The Scope of Usage-Based Theory. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00255

Im, H. S. (2018). Self-efficacy, Culture, and English Proficiency of University Students in South Korea. *Studies in English Language & Literature*, *44*(4), 111–129. https://doi.org/10.21559/AELLK.2018.44.4.006

Internet Archive. (n.d.). *OpenLibrary*. Retrieved 8 August 2022, from https://openlibrary.org/search/authors?q=

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1–11. https://doi.org/10.1016/j.jml.2017.09.002

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (Corrected at 8th printing). Springer. https://doi.org/10.1007/978-1-4614-7138-7

James, W. (1884). On Some Omissions of Introspective Psychology. *Mind*, *9*(33), 1–26.

Jeon, E. H., & Yamashita, J. (2014). L2 Reading Comprehension and Its Correlates: A Meta-Analysis: L2 Reading and Its Correlates. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Jeon, E. H., & Yamashita, J. (2022). L2 reading comprehension and its correlates: An updated meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Bilingual*

Processing and Acquisition (Vol. 13, pp. 29–86). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.03jeo

Jeong, Y. J., & Gweon, G. (2021). Advantages of Print Reading over Screen Reading: A Comparison of Visual Patterns, Reading Performance, and Reading Attitudes across Paper, Computers, and Tablets. *International Journal of Human–Computer Interaction*, *37*(17), 1674–1684. https://doi.org/10.1080/10447318.2021.1908668

Jian, Y.-C. (2022). Reading in print versus digital media uses different cognitive strategies: Evidence from eye movements during science-text reading. *Reading and Writing*, *35*(7), 1549–1568. https://doi.org/10.1007/s11145-021-10246-2

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, *73*(6), 841–855. https://doi.org/10.1177/1747021819897560

Johns, B. T., & Jamieson, R. K. (2018). A Large-Scale Analysis of Variance in Written Language. *Cognitive Science*, *42*(4), 1360–1374. https://doi.org/10.1111/cogs.12583

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2016). Experience as a Free Parameter in the Cognitive Modeling of Language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*(0). https://escholarship.org/uc/item/9ww798xm

Johnson, D. R. (2013). Transportation into literary fiction reduces prejudice against and increases empathy for Arab-Muslims. *Scientific Study of Literature*, *3*(1), 77–92. https://doi.org/10.1075/ssol.3.1.08joh

Johnson, M. (2018). The Embodiment of Language. In A. Newen, L. De Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 622–640). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.33

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. In *Psychology of Learning and Motivation* (Vol. 67, pp. 239–283). Elsevier. https://doi.org/10.1016/bs.plm.2017.03.008

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The Role of Semantic Diversity in Lexical Organization. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *66*(2), 115–124. https://doi.org/10.1037/a0026727

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?*. *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaufmann, S. (2017, May 9). The Best Way to Learn a New Language. *LingQ*. https://medium.com/the-linguist-on-language/the-best-way-to-learn-a-new-language-f1af92d756db

Kessler, R., & Beck, S. D. (2022). *L1 and L2 Learners Keep Their Eyes on the Prize: Eye-tracking Evidence during Idiom Recognition.* https://doi.org/10.15496/publikation-75926

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Kidd, D. C., & Castano, E. (2013). Reading Literary Fiction Improves Theory of Mind. *Science*, *342*(6156), 377–380. https://doi.org/10.1126/science.1239918

Kidd, D. C., & Castano, E. (2017). Different stories: How levels of familiarity with literary and genre fiction relate to mentalizing. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(4), 474–486. https://doi.org/10.1037/aca0000069

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kim, H., & Grüter, T. (2021). Predictive Processing of Implicit Causality in a Second Language: A Visual-World Eye-Tracking Study. *Studies in Second Language Acquisition*, *43*(1), 133–154. https://doi.org/10.1017/S0272263120000443

Kim, H., & Krashen, S. (1998). The author recognition and magazine recognition tests, and free voluntary reading as predictors of vocabulary development in English as a foreign language for Korean high school students. *System*, *26*(4), 515–523. https://doi.org/10.1016/S0346-251X(98)00035-9

Kolinsky, R., & Morais, J. (2018). The worries of wearing literate glasses. *L'Année Psychologique*, *Vol. 118*(4), 321–347. https://doi.org/10.3917/anpsy1.184.0321

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M., ... Usal, K. A. (2023). Text reading in English as a

second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, *45*(1), 3–37. https://doi.org/10.1017/S0272263121000954

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802–823. https://doi.org/10.1037/a0030859

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In M. Bar (Ed.), *Predictions in the Brain: Using our Past to Generate a Future* (pp. 190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.001.0001

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in Linear Mixed Effects Models* (Version 3.1-3) [Computer software]. https://CRAN.R-project.org/package=lmerTest

Laken, P. van der, & Lambert, L. (2023). *corrtable: Creates and Saves Out a Correlation Table with Significance Levels Indicated* (Version 0.1.1) [Computer software]. https://cran.r-project.org/web/packages/corrtable/index.html

Lakoff, G. (1990). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Lakoff, G. (2012). Explaining Embodied Cognition Results. *Topics in Cognitive Science*, *4*(4), 773–785. https://doi.org/10.1111/j.1756-8765.2012.01222.x

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford University Press.

Le Roux, B., & Rouanet, H. (2010a). *Multiple correspondence analysis*. Sage Publications.

Le Roux, B., & Rouanet, H. (2010b). The Method of Multiple Correspondence Analysis. In B. Le Roux & H. Rouanet, *Multiple correspondence analysis* (pp. 34–67). Sage Publications. https://doi.org/10.4135/9781412993906.d19

Lee, C. (2023). Language ideology and social class in Korean society. *The Journal of Linguistics Science*, *104*, 89–111. https://doi.org/10.21296/jls.2023.03.104.89

Lee, H., Seong, E., Choi, W., & Lowder, M. W. (2019). Development and assessment of the Korean Author Recognition Test. *Quarterly Journal of Experimental Psychology*, *72*(7), 1837–1846. https://doi.org/10.1177/1747021818814461

Lehtinen, N., Kautto, A., & Renvall, K. (2023). Frequent native language use supports phonemic and semantic verbal fluency in L1 and L2: An extended analysis of verbal fluency task performance in an L1 language attrition population. *International Journal of Bilingualism*, *28*(5), 884–906. https://doi.org/10.1177/13670069231193727

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition*, *36*(6), 1103–1121. https://doi.org/10.3758/MC.36.6.1103

Lin, M.-C., & Chih, Y.-C. (2023). Application of the Semantic Fluency Test in the Screening of Mandarin-Chinese-Speaking Older Adults with Mild Dementia of the Alzheimer Type. *Behavioral Sciences*, *13*(8), 635. https://doi.org/10.3390/bs13080635

Lüdecke, D., Bartel, A., Schwemmer, C., Powell, C., Djalovski, A., & Titz, J. (2023). *sjPlot: Data Visualization for Statistics in Social Science* (Version 2.8.15) [Computer software]. https://cran.r-project.org/web/packages/sjPlot/index.html

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, *114*(1), 29–41. https://doi.org/10.1016/j.cognition.2009.08.014

Macoir, J., Sylvestre, A., & Turgeon, Y. (2006). Classical Tests for Speech and Language Disorders. In *Encyclopedia of Language & Linguistics* (2nd ed., pp. 439–445). Elsevier. https://doi.org/10.1016/B0-08-044854-2/04191-2

Magno, C. (2010). Korean Students' Language Learning Strategies and Years of Studying English as Predictors of Proficiency in English. *TESOL Journal*, *2*, 39–61.

Mangen, A., Olivier, G., & Velay, J.-L. (2019). Comparing Comprehension of a Long Text Read in Print Book and on Kindle: Where in the Text and When in the Story? *Frontiers in Psychology*, *10*, 38. https://doi.org/10.3389/fpsyg.2019.00038

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68. https://doi.org/10.1016/j.ijer.2012.12.002

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Mano, Q. R., & Guerin, J. M. (2018). Direct and indirect effects of print exposure on silent reading fluency. *Reading and Writing*, *31*(2), 483–502. https://doi.org/10.1007/s11145-017-9794-5

Mar, R. A., Oatley, K., Hirsh, J., dela Paz, J., & Peterson, J. B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*, *40*(5), 19. https://doi.org/10.1016/j.jrp.2005.08.002

Mar, R. A., Oatley, K., & Peterson, J. B. (2009). Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes. *COMM*, *34*(4), 407–428. https://doi.org/10.1515/COMM.2009.025

Mar, R. A., & Rain, M. (2015). Narrative Fiction and Expository Nonfiction Differentially Predict Verbal Ability. *Scientific Studies of Reading*, *19*(6), 419–433. https://doi.org/10.1080/10888438.2015.1069296

Martin-Chang, S., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284. https://doi.org/10.1111/j.1467-9817.2008.00371.x

Martin-Chang, S., Kozak, S., & Rossi, M. (2020). Time to read Young Adult fiction: Print exposure and linguistic correlates in adolescents. *Reading and Writing*, *33*(3), 741–760. https://doi.org/10.1007/s11145-019-09987-y

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Masterson, J., & Hayes, M. (2007). Development and data for UK versions of an author and title recognition test for adults. *Journal of Research in Reading*, *30*(2), 212–219. https://doi.org/10.1111/j.1467-9817.2006.00320.x

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380. https://doi.org/10.3758/BF03206780

McCarron, S. P. (2026). Author Recognition Tests. In H. Nesi & P. Milin (Eds.), *Encyclopedia of Language & Linguistics* (3rd ed.). Elsevier. https://doi.org/10.1016/B978-0-323-95504-1.00497-X

McCarron, S. P., & Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods*, *53*(5), 2226–2237. https://doi.org/10.3758/s13428-021-01556-y

McCarron, S. P., & Kuperman, V. (2022). Effects of year of post-secondary study on reading skills for L1 and L2 speakers of English. *Journal of Research in Reading*, *45*(1), 43–64. https://doi.org/10.1111/1467-9817.12380

McCarron, S. P., Murphy, V. A., & Nation, K. (2025). An "Author Fluency Task": Semantic fluency as predictor of L2 vocabulary knowledge. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S136672892510045X

McClelland, C., & Powell, J. (2021, March 12). *Are people still reading physical books?* https://www.kantar.com/uki/inspiration/sport-leisure/are-people-still-reading-physical-books

McCreath, G. A., Linehan, C. M. J., & Mar, R. A. (2017). Can Differences in Word Frequency Explain Why Narrative Fiction Is a Better Predictor of Verbal Ability than Nonfiction? *Discourse Processes*, *54*(5–6), 373–381. https://doi.org/10.1080/0163853X.2017.1289794

McQuillan, J. (2019). Where Do We Get Our Academic Vocabulary? Comparing the Efficiency of Direct Instruction and Free Voluntary Reading. *The Reading Matrix*, *19*(1), 10.

McWhorter, J. (2008). *Our magnificent bastard tongue: The untold history of English*. Penguin Publishing Group.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*(1), 114–135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Menard, S. (2002). *Applied Logistic Regression Analysis*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983433

Mendelsund, P. (2014). *What we see when we read: A phenomenology with illustrations*. Vintage Books.

Merkle, E., You, D., Schneider, L., & Bae, S. (2020). *nonnest2: Tests of Non-Nested Models* (Version 0.5-5) [Computer software]. https://CRAN.R-project.org/package=nonnest2

Metz, C. (2024, October 23). Former OpenAI Researcher Says the Company Broke Copyright Law. *The New York Times*. https://www.nytimes.com/2024/10/23/technology/openai-copyright-law.html

Microsoft. (2024, February 17). *Microsoft Azure text to speech documentation*. Text to Speech Documentation. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-text-to-speech

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding. *Child Development*, *78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296. https://doi.org/10.1037/a0021890

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109. https://doi.org/10.3758/s13428-014-0534-3

Morais, J., & Kolinsky, R. (2021a). Seeing thought: A cultural cognitive tool. *Journal of Cultural Cognitive Science*, *5*(2), 181–228. https://doi.org/10.1007/s41809-020-00059-0

Morais, J., & Kolinsky, R. (2021b). Seeing thought in the future: Literate forecasting and forecasting literacy. *Journal of Cultural Cognitive Science*, *5*(2), 229–265. https://doi.org/10.1007/s41809-021-00085-6

Mori, Y., Kuroda, M., & Makino, N. (2016). Multiple Correspondence Analysis. In Y. Mori, M. Kuroda, & N. Makino, *Nonlinear Principal Component Analysis and Its Applications* (pp. 21–28). Springer Singapore. https://doi.org/10.1007/978-981-10-0159-8_3

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 109–120. https://doi.org/10.1037/aca0000089

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *Npj Science of Learning*, *2*(1), 3. https://doi.org/10.1038/s41539-017-0004-7

Nation, K., Dawson, N. J., & Hsiao, Y. (2022). 'Book language' and its implications for children's language, literacy, and development. *Current Directions in Psychological Science*, *31*(4), 375–380. https://doi.org/10.1177/09637214221103264

Nation, P. (2012). *The Vocabulary Size Test*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nattinger, J. R., & DeCarrico, J. S. (2010). *Lexical phrases and language teaching* (Nachdr.). Oxford Univ. Press.

Oakey, D. (2002). Formulaic language in English academic writing. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation*. J. Benjamins.

Oh, I. (2010). Education and Development: Why are Koreans Obsessed with Learning? *Comparative Sociology*, *9*(3), 308–327. https://doi.org/10.1163/156913209X12499527665422

Olsen, G. D., Pracejus, J. W., & O'Guinn, T. C. (2012). Print advertising: White space. *Journal of Business Research*, *65*(6), 855–860. https://doi.org/10.1016/j.jbusres.2011.01.007

OpenAI. (2024). *DALL·E 3* [Computer software]. https://chat.openai.com/

Organisation for Economic Co-operation and Development (OECD). (2010). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. OECD. https://doi.org/10.1787/9789264062658-en

Pae, H. K. (2011). Is Korean a syllabic alphabet or an alphabetic syllabary. *Writing Systems Research*, *3*(2), 103–115. https://doi.org/10.1093/wsr/wsr002

Pae, H. K. (2024a). Not Optimal yet Near-Optimal Writing System and Hangul. In H. K. Pae (Ed.), *Analyzing the Korean Alphabet: The Science of Hangul* (pp. 99–119). Springer International Publishing. https://doi.org/10.1007/978-3-031-49633-2_4

Pae, H. K. (2024b). The Topology of Hangul: Learnability, Efficiency, and Utility. In H. K. Pae (Ed.), *Analyzing the Korean Alphabet: The Science of Hangul* (pp. 187–202). Springer International Publishing. https://doi.org/10.1007/978-3-031-49633-2_7

Pae, H. K., Bae, S., & Yi, K. (2019). More than an alphabet: Linguistic features of Korean and their influences on Hangul word recognition. *Written Language and Literacy*, *22*(2), 223–246. https://doi.org/10.1075/wll.00027.pae

Papoutsaki, A. (2015). Scalable Webcam Eye Tracking by Learning from User Interactions. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 219–222. https://doi.org/10.1145/2702613.2702627

Papoutsaki, A., Daskalova, N., Sangkloy, P., Huang, J., Laskey, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 3839–3845.

Pelletier, J. (2006). Relations Among Theory of Mind, Metacognitive Language, Reading Skills and Story Comprehension In L1 and L2 Learners. In A. Antonietti, O. Liverta-Sempio, & A. Marchetti (Eds.), *Theory of Mind and Language in Developmental Contexts* (pp. 77–92). Springer US. https://doi.org/10.1007/0-387-24997-4_4

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research*, *21*(3), 381–402. https://doi.org/10.1177/1362168815618428

Pérez, A. I., Fotiadou, G., & Tsimpli, I. (2022). Preserved Executive Control in Ageing: The Role of Literacy Experience. *Brain Sciences*, *12*(10), 1392. https://doi.org/10.3390/brainsci12101392

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, *14*, 84–94. https://doi.org/10.1016/j.jeap.2014.01.002

Perfetti, C. (1985). *Reading ability*. Oxford University Press.

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy*. John Benjamins publishing company.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*(1), 75–94. https://doi.org/10.1177/1362168813505384

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113–138. https://doi.org/10.1177/1362168814568131

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Prolific. (n.d.). https://www.prolific.com/

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In Search of Matthew Effects in Reading. *Journal of Learning Disabilities*, *49*(5), 499–514. https://doi.org/10.1177/0022219414559974

Proverbio, A. M., Lilli, S., Semenza, C., & Zani, A. (2001). ERP indexes of functional differences in brain activation during proper and common names retrieval. *Neuropsychologia*, *39*(8), 815–827. https://doi.org/10.1016/S0028-3932(01)00003-3

Prystauka, Y., Altmann, G. T. M., & Rothman, J. (2023). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*, *56*(4), 3504–3522. https://doi.org/10.3758/s13428-023-02176-4

Puig-Mayenco, E., Chaouch-Orozco, A., Liu, H., & Martín-Villena, F. (2023). The LexTALE as a measure of L2 global proficiency: A cautionary tale based on

a partial replication of Lemhöfer and Broersma (2012). *Linguistic Approaches to Bilingualism*, *13*(3), 299–314. https://doi.org/10.1075/lab.22048.pui

Purpuri, S., Vasta, N., Filippi, R., Wei, L., & Mulatti, C. (2023). The Foreign Language Effect on Tolerance of Ambiguity. *Bilingualism: Language and Cognition*, 1–9. https://doi.org/10.1017/S1366728923000469

Qiu, M., Castro, N., & Johns, B. (2024). Estimating Type of Print Exposure across Aging through Author Production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/41z38291

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Restrepo-Ramos, F. D. (2015). Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. *PROFILE Issues in Teachers' Professional Development*, *17*(1), 157–166. https://doi.org/10.15446/profile.v17n1.43957

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.4.3) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Ricketts, J., Nation, K., & Bishop, D. V. M. (2007). Vocabulary Is Important for Some, but Not All Reading Skills. *Scientific Studies of Reading*, *11*(3), 235–257. https://doi.org/10.1080/10888430701344306

Ringbom, H. (2016). Comprehension, Learning and Production of Foreign Languages: The Role of Transfer. In R. Alonso Alonso (Ed.), *Crosslinguistic Influence in Second Language Acquisition* (pp. 38–52). Multilingual Matters. https://doi.org/10.21832/9781783094837

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Rodrigo, V., McQuillan, J., & Krashen, S. (1996). Free Voluntary Reading and Vocabulary Knowledge in Native Speakers of Spanish. *Perceptual and Motor Skills*, *83*(2), 648–650. https://doi.org/10.2466/pms.1996.83.2.648

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002

Roose, K. (2023, February 16). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Sage, K., Piazzini, M., Downey IV, J. C., & Masilela, L. (2020). Reading from print, laptop computer, and e-reader: Differences and similarities for college students' learning. *Journal of Research on Technology in Education*, *52*(4), 441–460. https://doi.org/10.1080/15391523.2020.1713264

Samur, D., Tops, M., & Koole, S. L. (2018). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion*, *32*(1), 130–144. https://doi.org/10.1080/02699931.2017.1279591

Saslow, M. G. (1967). Latency for Saccadic Eye Movement*. *Journal of the Optical Society of America*, *57*(8), 1030. https://doi.org/10.1364/JOSA.57.001030

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmitt, N. (2010). *Researching Vocabulary*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230293977

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Senaldi, M. S. G., Titone, D. A., & Johns, B. T. (2022). Determining the importance of frequency and contextual diversity in the lexical organization of multiword expressions. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *76*(2), 87–98. https://doi.org/10.1037/cep0000271

Sénéchal, M., LeFevre, J.-A., Hudson, E., & Lawson, E. P. (1996). Knowledge of storybooks as a predictor of young children's vocabulary. *Journal of*

Educational Psychology, *88*(3), 520–536. https://doi.org/10.1037/0022-0663.88.3.520

Serva, M., & Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, *81*(6), 68005. https://doi.org/10.1209/0295-5075/81/68005

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00772

Shatil, E., Share, D. L., & Levin, I. (2000). On the contribution of kindergarten writing to grade 1 literacy: A longitudinal study inHebrew. *Applied Psycholinguistics*, *21*(1), 1–21. https://doi.org/10.1017/S0142716400001016

Shin, D. (2007). The High Frequency Collocations of Spoken and Written English. *English Teaching*, *62*(1), 199–218. https://doi.org/10.15858/engtea.62.1.200703.199

Shin, J. C. (2012). Higher education development in Korea: Western university ideas, Confucian tradition, and economic development. *Higher Education*, *64*(1), 59–72. https://doi.org/10.1007/s10734-011-9480-5

Siddiqui, S., West, R. F., & Stanovich, K. E. (1998). The Influence of Print Exposure on Syllogistic Reasoning and Knowledge of Mental-State Verbs. *Scientific Studies of Reading*, *2*(1), 81–96. https://doi.org/10.1207/s1532799xssr0201_4

Siegelman, N., Elgort, I., Brysbaert, M., Agrawal, N., Amenta, S., Arsenijević Mijalković, J., Chang, C. S., Chernova, D., Chetail, F., Clarke, A. J. B., Content, A., Crepaldi, D., Davaabold, N., Delgersuren, S., Deutsch, A.,

Dibrova, V., Drieghe, D., Filipović Đurđević, D., Finch, B., ... Kuperman, V. (2024). Rethinking First Language–Second Language Similarities and Differences in English Proficiency: Insights From the ENglish Reading Online (ENRO) Project. *Language Learning*, *74*(1), 249–294. https://doi.org/10.1111/lang.12586

Sieghart, M. A. (2021, July 9). Why do so few men read books by women? *The Guardian*. https://www.theguardian.com/books/2021/jul/09/why-do-so-few-men-read-books-by-women

Sieghart, M. A. (2022). *The Authority Gap: Why women are still taken less seriously than men, and what we can do about it*. Black Swan.

Silberling, A. (2024, August 27). Why AI can't spell 'strawberry'. *TechCrunch*. https://techcrunch.com/2024/08/27/why-ai-cant-spell-strawberry/

Singer, L. M., & Alexander, P. A. (2017). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *The Journal of Experimental Education*, *85*(1), 155–172. https://doi.org/10.1080/00220973.2016.1143794

Siyanova, A., & Schmitt, N. (2008). L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *The Canadian Modern Language Review*, *64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*(2), 251–272. https://doi.org/10.1177/0267658310382068

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776–784. https://doi.org/10.1037/a0022531

Slim, M. S., & Hartsuiker, R. J. (2023). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js. *Behavior Research Methods*, *55*(7), 3786–3804. https://doi.org/10.3758/s13428-022-01989-z

Snow, C. E. (2010). Academic Language and the Challenge of Reading for Learning About Science. *Science*, *328*(5977), 450–452. https://doi.org/10.1126/science.1182597

Son, O., Ryu, S., & Yun, H. (2022). Developing and testing validity of LexTALE-Korean: An efficient tool for Korean proficiency assessment. *Language and Information*, *26*(2), 87–107. https://doi.org/10.29403/LI.26.2.4

Sonbul, S., & Schmitt, N. (2013). Explicit and Implicit Lexical Knowledge: Acquisition of Collocations Under Different Input Conditions. *Language Learning*, *63*(1), 121–159. https://doi.org/10.1111/j.1467-9922.2012.00730.x

Sparks, R. L. (1995). Examining the Linguistic Coding Differences Hypothesis to Explain Individual Differences in Foreign Language Learning. *Annals of Dyslexia*, *45*, 187–214.

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012). Do L1 Reading Achievement and L1 Print Exposure Contribute to the Prediction of L2

Proficiency? *Language Learning*, *62*(2), 473–505. https://doi.org/10.1111/j.1467-9922.2012.00694.x

Spear-Swerling, L., Brucker, P. O., & Alfano, M. P. (2010). Relationships between sixth-graders' reading comprehension and two different measures of print exposure. *Reading and Writing*, *23*(1), 73–96. https://doi.org/10.1007/s11145-008-9152-8

Stainthorp, R. (1997). A Children's Author Recognition Test: A Useful Tool in Reading Research. *Journal of Research in Reading*, *20*(2), 148–158. https://doi.org/10.1111/1467-9817.00027

Stanovich, K. E. (1986). Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, *20*(1), 51–68. https://doi.org/10.3758/BF03208254

Stanovich, K. E., & Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology*, *85*(2), 211–229. https://doi.org/10.1037/0022-0663.85.2.211

Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, *24*(4), 402–433. https://doi.org/10.2307/747605

Stoops, A., & Montag, J. L. (2023). Effects of individual differences in text exposure on sentence comprehension. *Scientific Reports*, *13*(1), 16812. https://doi.org/10.1038/s41598-023-43801-8

Storr, W. (2021). *The science of storytelling: Why stories make us human and how to tell them better*. Abrams Press.

Strømsø, H. I. (2023). Does students' exposure to websites moderate the positive relationship between print exposure and text comprehension? *Reading and Writing*, *37*(8), 2151–2171. https://doi.org/10.1007/s11145-023-10468-6

Su, Y., Li, Y., & Li, H. (2023). Development and validation of the simplified Chinese Author Recognition Test: Evidence from eye movements of Chinese adults in Mainland China. *Journal of Research in Reading*, *n/a*(n/a). https://doi.org/10.1111/1467-9817.12437

Summers, K. (2013). Adult Reading Habits and Preferences in Relation to Gender Differences. *Reference & User Services Quarterly*, *52*(3), 243–249.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 523–534. https://doi.org/10.1016/S0022-5371(79)90284-6

Szudarski, P. (2017). Learning and Teaching L2 Collocations: Insights from Research. *TESL Canada Journal*, *34*(3). https://doi.org/10.18806/tesl.v34i3.1280

Szudarski, P., & Conklin, K. (2014). Short- and Long-Term Effects of Rote Rehearsal on ESL Learners' Processing of L2 Collocations. *TESOL Quarterly*, *48*(4), 833–842. https://doi.org/10.1002/tesq.201

Tager-Flusberg, H. (2007). Evaluating the Theory-of-Mind Hypothesis of Autism. *Current Directions in Psychological Science*, *16*(6), 311–315. https://doi.org/10.1111/j.1467-8721.2007.00527.x

Taylor, H. (2022). *Why Women Read Fiction: The Stories of Our Lives*. Oxford University Press.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance. *The Journal of General Education*, *58*(3), 129–151. https://doi.org/10.2307/27798135

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209–253. https://doi.org/10.1016/S0010-0277(99)00069-4

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cogl*, *11*(1–2), 61–82. https://doi.org/10.1515/cogl.2001.012

Tomasello, M. (Ed.). (2010). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative Data Stratified by Age and Education for Two Measures of Verbal Fluency: FAS and Animal Naming. *Archives of Clinical Neuropsychology*, *14*(2), 167–177. https://doi.org/10.1016/S0887-6177(97)00095-4

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138–146. https://doi.org/10.1037/0894-4105.11.1.138

Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society*, *4*, 137–143. https://doi.org/10.1017/s1355617798001374

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Tskhovrebova, E., Zufferey, S., & Gygax, P. (2022). Individual Variations in the Mastery of Discourse Connectives from Teenage Years to Adulthood. *Language Learning*, *72*, 412–455. https://doi.org/10.1111/lang.12481

Tunmer, W. E., Chapman, J., & Prochnow, J. (2006). Literate cultural capital at school entry predicts later reading achievement: A seven year longitudinal study. *New Zealand Journal of Educational Studies*, *41*, 183–204.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.

Tyler, A. (2010). Usage-Based Approaches to Language and Their Applications to Second Language Learning. *Annual Review of Applied Linguistics*, *30*, 270–291. https://doi.org/10.1017/S0267190510000140

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *Quarterly Journal of Experimental Psychology*, *64*(3), 447–466. https://doi.org/10.1080/17470218.2010.505292

van Bergen, E., Snowling, M. J., Zeeuw, E. L., Beijsterveldt, C. E. M., Dolan, C. V., & Boomsma, D. I. (2018). Why do children read more? The influence of

reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, *59*(11), 1205–1214. https://doi.org/10.1111/jcpp.12910

Van Silfhout, G., Evers-Vermeul, J., & Sanders, T. (2015). Connectives as Processing Signals: How Students Benefit in Processing Narrative and Expository Texts. *Discourse Processes*, *52*(1), 47–76. https://doi.org/10.1080/0163853X.2014.905237

Vander Beken, H., & Brysbaert, M. (2018). Studying texts in a second language: The importance of test type. *Bilingualism: Language and Cognition*, *21*(5), 1062–1074. https://doi.org/10.1017/S1366728917000189

Vander Beken, H., De Bruyne, E., & Brysbaert, M. (2020). Studying texts in a non-native language: A further investigation of factors involved in the L2 recall cost. *Quarterly Journal of Experimental Psychology*, *73*(6), 891–907. https://doi.org/10.1177/1747021820910694

Verhoeven, L. T. (1994). Transfer in Bilingual Development: The Linguistic Interdependence Hypothesis Revisited. *Language Learning*, *44*(3), 381–415. https://doi.org/10.1111/j.1467-1770.1994.tb01112.x

Vermeiren, H., & Brysbaert, M. (2023). How useful are native language tests for research with advanced second language users? *Bilingualism: Language and Cognition*, *27*(1), 204–213. https://doi.org/10.1017/S1366728923000421

Vermeiren, H., Vandendaele, A., & Brysbaert, M. (2022). Validated tests for language research with university students whose native language is English: Tests of vocabulary, general knowledge, author recognition, and reading comprehension. *Behavior Research Methods*, *55*(3), 1036–1068. https://doi.org/10.3758/s13428-022-01856-x

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2005). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models* (pp. xv, 340). Springer Publishing Co.

Vu, D. V., & Peters, E. (2023). A Longitudinal Study on the Effect of Mode of Reading on Incidental Collocation Learning and Predictors of Learning Gains. *TESOL Quarterly*, *57*(1), 5–32. https://doi.org/10.1002/tesq.3111

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, *57*(2), 307. https://doi.org/10.2307/1912557

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

Webb, S., & Kagimoto, E. (2009). The Effects of Vocabulary Learning on Collocation and Meaning. *TESOL Quarterly*, *43*(1), 55–77. https://doi.org/10.1002/j.1545-7249.2009.tb00227.x

Webb, S., Newton, J., & Chang, A. (2013). Incidental Learning of Collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

West, R. F., Stanovich, K. E., & Mitchell, H. R. (1993). Reading in the Real World and Its Correlates. *Reading Research Quarterly*, *28*(1), 35–50. https://doi.org/10.2307/747815

Wetzel, M., Zufferey, S., & Gygax, P. (2020). Second Language Acquisition and the Mastery of Discourse Connectives: Assessing the Factors That Hinder L2-Learners from Mastering French Connectives. *Languages*, *5*(3), 35. https://doi.org/10.3390/languages5030035

White, L. (1990). Second language acquisition and universal grammar. *Studies in Second Language Acquisition*, *12*(2), 121–133.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Brand, T. van den, Posit, & PBC. (2024). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (Version 3.5.1) [Computer software]. https://cran.r-project.org/web/packages/ggplot2/index.html

Williams, D. (2018). Predictive Processing and the Representation Wars. *Minds and Machines*, *28*(1), 141–172. https://doi.org/10.1007/s11023-017-9441-6

Willingham, D. T. (2014, August 8). *Ask the Cognitive Scientist: The Privileged Status of Story.* American Federation of Teachers. https://www.aft.org/periodical/american-educator/summer-2004/ask-cognitive-scientist

Wimmer, L., Currie, G., Friend, S., & Ferguson, H. J. (2022). The effects of reading narrative fiction on social and moral cognition: Two experiments following a multi-method approach. *Scientific Study of Literature.* https://doi.org/10.1075/ssol.21010.wim

Wimmer, L., & Ferguson, H. J. (2022). Testing the validity of a self-report scale, author recognition test, and book counting as measures of lifetime exposure to print fiction. *Behavior Research Methods.* https://doi.org/10.3758/s13428-021-01784-2

Woolf, V. (1937, April 29). *Craftsmanship* [Audio recording]. [Radio broadcast episode]. In *Words Fail Me.* BBC. https://www.bbc.co.uk/news/av/entertainment-arts-28231055

Woolf, V. (1942). *The Death of the Moth: And Other Essays*. New York, Harcourt, Brace and Company.

Woolf, V. (2012). *The Death of the Moth, and Other Essays*. Project Gutenberg. https://gutenberg.net.au/ebooks12/1203811h.html

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Wray, A. (2006). Formulaic Language. In *Encyclopedia of Language & Linguistics* (pp. 590–597). Elsevier. https://doi.org/10.1016/B0-08-044854-2/04777-5

Wulff, S. (2021). Usage-based approaches. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora*. Routledge.

Yamashita, J. (2013). Effects of extensive reading on reading attitudes in a foreign language. *Reading in a Foreign Language*, *25*(2), 248–263.

Zufferey, S., & Gygax, P. (2020). "Roger Broke His Tooth. *However*, He Went to the Dentist": Why Some Readers Struggle to Evaluate Wrong (and Right) Uses of Connectives. *Discourse Processes*, *57*(2), 184–200. https://doi.org/10.1080/0163853X.2019.1607446

Zufferey, S., & Gygax, P. M. (2017). Processing Connectives with a Complex Form-Function Mapping in L2: The Case of French "En Effet". *Frontiers in Psychology*, *8*(1198), 1–11. https://doi.org/10.3389/fpsyg.2017.01198

Zufferey, S., Mak, W., Degand, L., & Sanders, T. (2015). Advanced learners' comprehension of discourse connectives: The role of L1 transfer across on-line and off-line tasks. *Second Language Research*, *31*(3), 389–411. https://doi.org/10.1177/0267658315573349

Appendices

Appendix A: Measures

Appendix A.1 Author Fluency Task (AFT) Instructions

English Instructions:

Author Naming

English version

1) In this task, you will be given **3 minutes** to type in as many **names of authors** as you can recall **off the top of your head.**

Please only use names you are certain belong to an **author who has been published in English.** Both fiction and non-fiction writers are acceptable. For

example, you may include names of novelists and playwrights, as well as investigative journalists and historians.

You may use the name of any author from any time period or literary tradition, and who has written in any language, as long as the author has been published in English.

You do not need to have read these authors' books yourself—in fact, you don't even need to **like** the works of these authors—they only need to be authors who come to mind.

2) Please **do not look up names** of authors on Google or any other online resource. Similarly, please avoid looking around at any nearby bookshelves or magazines. Try to include both the author's first name (or initials) if possible, and surname. Don't worry if the spelling is not exact.

3) Please note that **authors must be known primarily for their written work.** This means that **public individuals** who have published books such as autobiographies but **who are not primarily known for their writing** (for example politicians, celebrities, comedians, etc.) **are excluded.**

For example, although Barack Obama has written multiple books, it may be unlikely that these books would be as widely-known were it not for his former position as U.S. president; for this reason, he would not be accepted as an author name.

4) Names of authors must be submitted by pressing **"Enter"** on Windows/PC or **"Return"** on a Mac. Once **3 minutes** have elapsed, your results will be saved, and you will not be able to modify them.

Once you have read through these instructions carefully, please click **"Begin"** to continue.

French Instructions:

Tâche d'énumération des auteurs

Version française

1) Dans cette tâche, vous aurez **trois minutes** pour taper autant de **noms d'auteurs** que vous pouvez vous rappeler **au pied levé.**

N'utilisez que les noms des **auteurs dont les œuvres ont été publiées en français.** Les auteurs de fictions et de littérature non romanesque sont tous les deux acceptables. Par exemple, vous pouvez inclure des noms des romanciers et des dramaturges, ainsi que des noms des journalistes d'enquête et des historiens.

Vous pouvez utiliser le nom de tout auteur de toute période ou tradition littéraire, à condition que celui-ci ait été publié en français.

2) **Il n'est pas nécessaire que vous ayez lu les livres de ces auteurs vous-même** — en effet, vous n'avez même pas besoin **d'aimer** les œuvres de ces auteurs — ils ont seulement besoin d'être les auteurs qui viennent à l'esprit.

Ne cherchez pas les noms des auteurs sur Google ou d'autres ressources en ligne. Parallèlement, veuillez éviter de regarder des étagères des bibliothèques ou les magazines à proximité. Essayez d'inclure le prénom (ou les initiales) et le nom de l'auteur si possible. Ne vous inquiétez pas si l'orthographe n'est pas exacte.

3) Veuillez noter **que les auteurs doivent être connus principalement pour leur œuvre écrite.** Ainsi, **des personnes publiques** qui ont publié des livres comme les autobiographies mais **qui ne sont pas principalement connus pour leur écriture** (par exemple, les politiciens, les célébrités, les comédiens, etc.) **sont exclues.**

Par exemple, même si Emmanuel Macron a écrit un livre, il est peu probable que ce livre soit aussi largement connu sans son poste en tant que président de la République. C'est pourquoi il ne serait pas accepté comme nom d'auteur.

4) Des noms d'auteurs doivent être soumis en appuyant sur la touche « Entrée » sur Windows/PC ou la touche « Retour » sur un Mac. Une fois que **trois minutes** se seront écoulées, vos résultats seront enregistrés et vous ne pourrez plus les modifier.

Une fois que vous aurez lu attentivement ces instructions, veuillez cliquer sur « **Commencer** » pour continuer.

Korean Instructions:

작가 이름 쓰기

한글판

1) 이 시험에서는 **3분간 기억할 수 있는 작가 이름**을 최대한 많이 입력하시면 됩니다.

한국어로 출판된 책의 작가 이름만 입력해주세요. 소설 및 논픽션 작가 모두 가능합니다. 예를 들어, 소설가 및 극작가, 탐사보도 기자 및 역사가의 이름을 포함할 수 있습니다.

사용자는 작가가 사용한 언어와 상관없이 한국어로 출판된 한 모든 시대와 문학적 전통의 모든 작가의 이름을 작성 가능합니다.

이 작가들의 책을 직접 읽어보았을 필요는 없습니다. 이 작가들의 작품을 **좋아할** 필요도 없습니다. 머릿속에 떠오르는 대로 작성해주시면 됩니다.

2) 구글이나 기타 온라인 자료에서 작가의 이름을 **검색하지 마세요.** 마찬가지로 가까운 책장이나 잡지도 둘러보지 마세요. 작성자의 이름(또는 이니셜)과 가능하면 성을 모두 포함하세요. 철자가 정확하지 않더라도 걱정하지 마세요.

3) **작가는 주로 본인의 출간물로 알려져 있어야 합니다.** 이는 자서전과 같은 책을 출간했지만 **본인의 출간물로 인해 주로 알려지지 않은 공인들은** (예를 들어 정치인, 연예인, 코미디언 등) **제외됩니다.**

예를 들어, 버락 오바마가 여러 권의 책을 썼지만, 이전 미국 대통령으로서의 지위가 없었다면 이 책들이 널리 알려지지 않았을 확률이 높습니다. 이러한 이유로, 그는 작가의 이름으로 받아들여지지 않습니다.

4) 작가 이름은 Windows/PC에서 "**Enter**"를 누르거나 Mac에서 "**Return**"를 눌러 입력해야 하며, **3분**이 경과하면 결과가 저장되어 수정할 수 없습니다.

설명들을 주의 깊게 읽으셨다면 "**시작**"을 눌러서 진행하세요.

Appendix A.2 Lexical decision task for fiction keywords

List of Stimuli (words/non-words) used in the Lexical Decision Task for fiction keywords.

Key - Words are in **Bold**, Non-words are *Italicised.*

lereknoreman	*uncery*	*wiltilus*
gumstoo	*ecicose*	*fecalpic*
prepossess	*lickem*	**disquiet**
ignominiously	*balicofos*	*atcarant*
hophound	*lamnomancer*	**epicure**
genteel	**resplendent**	**lichen**
lackey	**thrall**	**hangar**
gumshoe	*coccer*	*poque*
hingsionary	*rephrastent*	*inveggenicamo*
functionary	**halitosis**	**venturesome**
misspine	**smoky**	**defrock**
clebeood	**feckless**	**bawdy**
hothouse	*threll*	**denizen**
quogly	**clamber**	**pique**
luridly	*sunipen*	*dedrack*
hesslat	*rurry*	**ostentation**

hellcat

exult

blistry

rummy

unexceptionable

cauly

estimable

snugly

imbecility

madcap

iprilodiously

aspirant

nautilus

longshoreman

tuckless

beripe

heraldic

plebeian

derision

madtop

futter

busteil

gendiel

misspent

gristly

incommunicado

bereft

ornery

fibren

malady

gluky

hackneyed

hancor

ventumpmime

coffer

autibible

denwriet

emuor

illefarity

fuledly

necromancer

totter

spimber

orogmoptionable

uttincation

prepirress

bestial

sunady

Appendix A.3 Reading Habits Questionnaires

Reading Habits Questionnaire v1 (L1, Chapter 2)

The outcome variable for this measure was a composite score comprised of the following self-report questions about individual reading habits.

1) Reading for pleasure: This was measured using responses to the question, *"Approximately how much time do you spend reading for pleasure (i.e. not for work or a school course) each week?"* The following 7 ordinal response ranges were available to select: *"None, Less than 1 hour, 1-3 hours, 3-5 hours, 5-10 hours, 10-20 hours, 20+ hours"*

2) Number of books in the home: This was measured using responses to the question, *"About how many physical books were there in your home when you were 16 years old? Do not include magazines, newspapers or schoolbooks. (To give an estimation, one metre of shelving is about 40 books.)"* The following 6 ordinal response ranges were available to select: *"10 books or fewer, 11-25 books, 26-100 books, 101-200 books, 201-500 books, 500+ books".* This question was adapted slightly from the PIAAC survey (Organisation for Economic Co-operation and Development (OECD), 2010).

3) Number of books read in the past year: This was measured using responses to the question, *"Not including textbooks, approximately how many books (or e-Books) have you read in the past year?"* The following 9 ordinal response ranges were available to select: *"None, 1-5, 6-10, 11-20, 21-30, 31-40, 41-*

50, 50-75, 76+".

4) Time spent reading news: This will be measured using responses to the question, *"Approximately how much time each day do you spend reading news articles (online or newspaper)?"* The following 6 ordinal response ranges were available to select: *"None, 1-15 minutes, 15-30 minutes, 30-60 minutes, 1-2 hours, 2+ hours."*

Items 1-4 were summed to form a composite reading habits score. For example, consider a hypothetical participant with the following response pattern:

1) 2 (Less than 1 hour)
2) 2 (11-25 books)
3) 4 (11-20 books)
4) 3 (15-30 minutes)

These items would be summed (2+2+4+3) to the composite score, 11. The maximum score is the sum of the individual measures (7+6+9+6 = 28).

Reading Habits Revised (L1, Chapter 2)

This revised questionnaire replaced questions 2 and 4 from the previous list with the following items:

1) Fiction reading amount: This was measured using responses to the question, *"I read **fiction** (novels, etc.)..."* The following 7 ordinal response ranges were available to select: *"Never, Occasionally, Roughly Once a Month, Roughly Once a Week, Roughly Every Day".*

2) Digital vs. print reading percentage: This was measured using responses to the question, *"Approximately how often do you read books on **digital formats** (tablet, eReader, etc.), **as opposed to print** (traditional paper books)?"* The following 7 ordinal response ranges were available to select: *"Never (I only read traditional paper books), Rarely, Very Rarely, Sometimes, Often, Very Often, Always (I only read digital books)".*

The outcome variable for this revised measure was calculated in the same way as the previous version, with ordinal responses given numeric values and summed.

L2 Demographic and Reading Habits Questions

(1) How good is the quality of your writing in English?

Very Poor, Poor, Neutral, Good, Very Good

(2) What is your reading speed in English?

Very Slow, Slow, Average, Fast, Very Fast

(3) How good is your reading comprehension?

Very Poor, Poor, Neutral, Good, Very Good

(4) How many hours a day do you spend reading and writing?

0-0.5 hours 1-2 hours 2-3 hours 3-4 hours 4+ hours

(5) What percentage of this time do you spend reading or writing texts on social media websites (0-100)?

Appendix A.4 Motivation Survey

The 10-item Student Opinion Survey (Finney et al., 2016; Thelk et al., 2009) was used to assess motivation to complete the tasks. The measured variable was the average response for each question (maximum 5). A 5-item scale (1=Strongly Disagree, 5=Strongly Agree) is used to evaluate each statement. Negatively formulated statements (3, 4, 7, 9) are reverse-coded prior to scoring (per Thelk et al., 2009).

Instructions to Participants

> *Please think about all of the tests that you just completed. Mark the answer that best represents how you feel about each of the statements below.*

Motivation Survey

Key: ***Reverse Scored***

Table A.1: Motivation survey ("Student Opinion Survey"; Finney et al., 2016; Thelk et al., 2009).

Number	Question	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	Doing well on this test was important to me.	1	2	3	4	5
2	I engaged in good effort throughout this test.	1	2	3	4	5
3	***I am not curious about how I did on this test relative to others.***	1	2	3	4	5
4	***I am not concerned about the score I receive on this test.***	1	2	3	4	5
5	This was an important test to me.	1	2	3	4	5
6	I gave my best effort on this test.	1	2	3	4	5
7	***While taking this test, I could have worked harder on it.***	1	2	3	4	5
8	I would like to know how well I did on this test.	1	2	3	4	5

9	***I did not give this test my full attention while completing it.***	1	2	3	4	5
10	While taking this test, I was able to persist to completion of the task.	1	2	3	4	5

Appendix A.5 Media Use Questionnaire

This variable is a composite score comprised of responses to a set of three self-report questions. The maximum score is the sum of the individual measures (7+6+8), 21. The items are as follows:

1) **Time on social media:** This was measured using responses to the question, *"Approximately how much time each day do you spend on social media?"* The following 7 ordinal response ranges were available for participants to select: *"None, 1-15 minutes, 15-30 minutes, 30-60 minutes, 1-2 hours, 2-4 hours, 4+ hours."*

2) **Social media reading:** This was measured using responses to the question, *"Of this time spent on social media, approximately what proportion is spent reading text (e.g. Facebook/Twitter/blog posts) vs. watching videos?"* The following 6 ordinal response ranges were available for participants to select: *"None, 1-15%, 16-30%, 31-50%, 51-75%, 76-100%"*

3) **Time watching video media:** This will be measured using responses to the question, *"Approximately how much time each day do you spend watching television, streaming video, or playing video games?"* The following 8 ordinal response ranges will be available to participants to select: *"None (I don't consume any video media), 0-30 minutes, 30-60 minutes, 1-2 hours, 2-4 hours, 4-6 hours, 6-8 hours, 8+ hours".*

Item 2 is reverse coded prior to scoring, then items 1-3 are summed to calculate the composite score. For example, consider a participant with the following response pattern:

1) 6 (2-4 hours)
2) 5 (51-75%) (Converted score of 2)
3) 4 (1-2 hours)

Item 2 would first be reverse coded, such that the 5 (second-from-highest on the scale) would become a 2 (second-from-lowest). The scores of 6, 2, and 4 would then be summed to the composite score, 12.

Appendix A.6 English Author Recognition Test

This version of ART is from Vermeiren et al., (2022).

Key - Author Names are in **Bold.**

Daniel Acheson

John Grisham

John Punnett

John Green

Jacqueline Wilson

Dan Brown

Ernest Hemingway

Sara Lakin

Vladimir Nabokov

Harper Lee

Hilary Mantel

Kurt Vonnegut

Thomas Wolfe

Cassandra Clare

Rainbow Rowell

Jeff Kinney

Samuel Beckett

F. Scott Fitzgerald

Anthony Horowitz

Roald Dahl

Tom Clancy

Tim Singler

J.D. Salinger

David Walliams

James Joyce

Robert Teesdale

Stephanie Meyer

Kate Dicamillo

Khaled Hosseini

E. Buxton

John Kestley

Mahmoud Abdellah

George Orwell

Mark Robin

Suzanne Collins

Lemony Snickett

Danielle Steel

Karin Slaughter

Justine Wells

Kathryn Lightner

E.L. Wilford

Rick Riordan

Toni Morrison

Maya Angelou

Lisa Wingate

E.L. James

Isabel Allende

Michael Morpurgo

Jorge Eudoro Remache

Kelly Weaver

Haruki Murakami

Jane Jessup

Holly Black

Sarah J. Maas

Melanie Marrero Morales

Stephen King

J.R.R. Tolkien

Isaac Asimov

Veronica Roth

Richard Grigley

Gabriel Garcia Marquez

E.B. White

Agatha Christie

Jack London

James Patterson

Roberto Borsani

Terry Pratchett	Chiara Ricci	Kim Wassing	**Virginia Woolf**
Kyra Appels	**Ray Bradbury**	Roy Leeman	Yasushi Sugawara
Kettil Christoffersen	**J.K. Rowling**	Judith L. Schecter	**William Faulkner**
Ludwig Lorenz	**Margaret Atwood**	Peter Mitchell	Gabriel Just
T.S. Eliot	**Lucinda Riley**	Elizabeth Wigelsworth	**Ayn Rand**
Angie Thomas	**Ralph Ellison**	**Paulo Coelho**	
	Theresa Ziegler		

Appendix A.7 French Author Recognition Test

This French version of ART is from Zufferey and Gygax (2020).

Key - Author Names are in **Bold.**

Ken Follett	**Joyce Carol Oates**	**Gilles Legardinier**
Christina Johnson	François Robin	**Tom Clancy**
Paul Jaquet	**Patrick Modiano**	**Virginie Grimaldi**
Jay Peter Holmes	**Vladimir Nabokov**	Marius Garnier
Claude Simon	Pierre Nicolas	**Bernard Weber**
Fred Vargas	**Saint-John Perse**	David Harper Townsend

Keith Cartwright

Katherine Kreutz

Francine Jaquet

Sylvain Granier

Adán Guillén Zepeda

Jessica Ann Lewis

Maxime Chattam

Angelo Lucchese

Toni Morrison

Maryann Phillips

Henri Garcia

Ray Bradbury

Christiane Bonnet

Patricia Cornwell

Delphine de Vigan

Marc Levy

Jean-Paul Sartre

Christian Wilbrecht

Agnès Ledig

Amélie Nothomb

Édouard Louis

Isabel Allende

Jean-Christophe Grangé

Harry Coltheart

Danielle Steel

Alex D. Lemaire

Umberto Eco

Joël Dicker

Günter Grass

Miguel Ángel Asturias

Manuel Flores

Gabrielle Ricaud

Alice Maze

William Golding

Lincoln Woodward

Jean-Christophe Rufin

Emma Svensson

Arnaldur Indridson

Gérard De Villiers

Annie Leimkühler

Raymond Chandler

T. C. Boyle

Antoine Faure

Agathe Lefort

E. L. James

Martin Suter

Fabrice Roussel

Alan Tsing

David Ashley

Gabriel García Márquez

Guillaume Musso

A. C. Kelly

Joachim Duval

Louis Delpierre

Jonathan Björk

Valérie Perrin

Jakob Burger

Valérie Coupé

Alexandre Soljenitsyne

Jean Perrin

Patrick Banville

Arturo Garcia Pérez

Appendix A.8 Korean Author Recognition Test

This version of the Korean Author Recognition Test is from Lee et al. (2019).

Key - Author Names are in **Bold.**

아부라메 시노	스콧 알렉산더	**알베르 카뮈**	**채사장**
조정래	프레드릭 먼도우	전애선	**박훈**
한강	에드몽 드 포르탕	산들	카니얼 휘트먼
니노미야 카즈나리	홍희라	글로리아 맥컴버	**기욤 뮈소**
호메로스	천솔뫼	**오쿠다 히데오**	손영지
요한 볼프강 폰 괴테	**김진명**	**마이클 샌델**	**백석**
이설민	프레아 뮐레르	하나자와 루이	**신영복**
요나스 요나손	피터 플래거티	**에쿠니 가오리**	안나 스팅
헤르만 헤세	리사 우드워드	**김소월**	**미셸 푸코**
김난도	**김영하**	웨인 필버그	이설태
애덤 스미스	**베르나르 베르베르**	**혜민 스님**	**윤동주**
		해리 코튼허트	**나태주**

하나타와 이치로	**장 지글러**	윌터 도리슨	은서희
진목	**히가시노 게이고**	더크워스 글램	미뉴얼 스미스
칼 세이건	**정유정**	알피노 르베유르	히로유키 오시타
김희설	설민지	**프란츠 카프카**	**어니스트 헤밍웨이**
알랭 드 보통	**제인 오스틴**	**법륜**	장순태
J.R.R. 톨킨	권순보	**무라카미 하루키**	윤이석
리처드 도킨스	**김미경**	에우로스	정희열
신경숙	진소미	백구영	**더글라스 케네디**
스테판 허스턴	아투로 가르시아 페레즈		

Appendix A.9 LexTALE (English)

The English LexTALE is taken from Lemhöfer & Broersma (2012).

List of Words Used in the LexTALE (English)

Key - Words are in **Bold**, Fillers are *Italicised.*

platery	spaunch	magrity
denial	**allied**	**nourishment**

generic	**slain**	abergy
mensible	**recipient**	proom
scornful	exprate	**turmoil**
stoutly	**eloquence**	**carbohydrate**
ablaze	**cleanliness**	**scholar**
kermshaw	**dispatch**	**turtle**
moonlit	rebondicate	fellick
lofty	**ingenious**	destription
hurricane	**bewitch**	**cylinder**
flaw	skave	**censorship**
alberation	**plaintively**	**celestial**
unkempt	kilp	**rascal**
breeding	interfate	purrage
festivity	**hasty**	pulsh
screech	**lengthy**	**muddy**
savory	**fray**	quirty
plaudate	crumper	pudour
shin	**upkeep**	**listless**
fluid	**majestic**	**wrought**

Appendix A.10 LexTALE (French)

The French LexTALE is taken from Brysbaert (2013).

Key - Words are in **Bold**.

pourcine	**pouce**	plaiser
fouet	**semonce**	**hache**
chameau	sentuelle	**peigne**
endifier	jamain	détume
cerveler	**lézard**	**ennemi**
fourmi	**marteau**	prioche
cadenas	**cloche**	**éventail**
cintre	**salière**	**abêtir**
pinceau	**amadouer**	**amorce**
caddie	**occire**	agire
parchance	**infâme**	**citrouille**
bouton	**panier**	**balai**
metter	**lanière**	**nouer**
treillage	**fenêtre**	procoreux
sacher	**esquif**	**robinet**
replaner	**remporter**	réporce
bouilloire	**tanin**	cessure

mignon	**église**	**indicible**
vicelard	**clouer**	parir
retruire	**poisson**	**fascine**
écouce	**oeillet**	**escroc**
orgueil	**fosse**	**écureuil**
crayon	joueux	osseaux
capeline	**inciter**	boutard
racaille	oeiller	rejoute
huif	soumon	**canoter**
gloque	**alourdir**	**cheveux**
honteur	**mappemonde**	**dauphin**

Appendix A.11 LexTALE (Korean)

The Korean LexTALE is taken from Son et al. (2022).

Key – Words are in **Bold.**

사람	**혐의**	**새롭다**	잔끅
민족	**기쁨**	**숱하다**	다프다
관련	**정착**	**함께**	별노
변화	**다짐**	**분명**	물곤
존재	**왜곡**	**특히**	디로
판매	**여백**	**흡사**	횡병
까닭	**증략**	**아늑히**	과더
파괴	**출몰**	국밀	노알
권리	**제각각**	다늘	단붕
바깥	**젖니**	완벙	곤계
논란	**등극**	바남	꼬링
또래	**알다**	예숙	마츰
엉터리	**나오다**	다지막	통억
현미경	**살다**	종됴	목조리
누더기	**모르다**	규노	신덩서
결국	**생기다**	문위기	모징

개발	**닿다**	소묵	부히
필요	**꿰차다**	후민	잭다
얼굴	**주다**	포잔	하매다
교육	**찾다**	호떨	뭅다
가운데	**부르다**	빌통	씬다
표현	**잃다**	기촌	거늬
과학	**굶다**	창다	같티
회담	**머금다**	단지다	
원칙	**없다**	남히	

Appendix A.12 Connectives Task

This task was translated to English from French, and was adapted where necessary from Wetzel et al. (2020). See

Table A.2 for comparison with original French sentences and complexity levels for each coherence relation.

Relation: Addition

Furthermore (Low Frequency)

Philip has many good qualities—he really likes to cook, _____ he's a good pianist.

a) therefore b) **furthermore** c) nevertheless d) whereas e) provided f) since

Marie and Emily have much in common—they both love gardening, _____ they love sailing.

a) despite this b) whereas c) hence d) as long as e) **furthermore** f) given that

Camilla is a real film fanatic—she likes to go to the cinema, _____ she has a large collection of old movie posters.

a) **furthermore** b) provided c) conversely d) despite this e) since f) therefore

Martin speaks very loudly into the telephone, _____ he never stops talking.

a) as long as b) hence c) despite this d) conversely e) **furthermore** f) given that

Leon seems very cultured—he listens to a lot of classical music, _____ he loves traditional dance.

a) given that b) as long as c) hence d) **furthermore** e) nevertheless f) whereas

Indeed (High Frequency)

Gabriel often recites poetry, _____ he was practicing a new poem last night.

a) **indeed** b) despite this c) whereas d) since e) therefore f) as long as

Arthur is always losing his things, _____ he misplaced his keys this morning.

a) provided b) conversely c) **indeed** d) hence e) despite this f) since

Juliet is always ready to help, _____ she volunteers regularly.

a) conversely b) **indeed** c) despite this d) given that e) hence f) provided

Lucas has made a few enemies, _____ he received a threatening letter recently.

a) **indeed** b) given that c) therefore d) as long as e) conversely f) nevertheless

Emma only drinks black tea, _____ it's her favourite beverage.

a) given that b) whereas c) provided d) hence e) nevertheless f) **indeed**

Relation: Consequence

Hence (Low Frequency)

David was often late when he first started, _____ he always sets his alarm a bit early each morning.

a) **hence** b) nevertheless c) conversely d) provided e) since f) furthermore

John doesn't have time to eat a full meal at work, _____ he only takes snacks to the office.

a) provided b) conversely c) indeed d) **hence** e) despite this f) given that

Max often forgets his glasses at home, _____ he keeps another pair at work.

a) whereas b) given that c) despite this d) as long as e) indeed f) **hence**

Jane took private lessons this term, _____ she did well in her exams.

a) whereas b) furthermore c) provided d) despite this e) given that f) **hence**

Mary is pregnant, _____ she is not drinking alcohol.

a) nevertheless b) as long as c) since d) **hence** e) indeed f) conversely

Therefore (High Frequency)

Martin avoids doing his mathematics homework, _____ he will struggle on the exam.

a) nevertheless b) whereas c) provided d) since e) furthermore f) **therefore**

Jane ran fast so as not to miss the train, _____ she was on time for work.

a) given that b) despite this c) furthermore d) **therefore** e) conversely f) as long as

Philip lives next to a grocery store, _____ he has everything he needs to cook.

a) whereas b) provided c) furthermore d) despite this e) **therefore** f) since

Nadia often goes running, _____ she's very fit.

a) given that b) nevertheless c) **therefore** d) indeed e) conversely f) as long as

Louise works in another city, _____ she takes the train every day.

a) as long as b) furthermore c) **therefore** d) whereas e) nevertheless f) since

Relation: Concession

Nevertheless (Low Frequency)

Kenneth does not like sports, _____ he does them to stay healthy.

a) furthermore b) **nevertheless** c) whereas d) therefore e) provided f) since

Joanne did her homework very well, _____ she couldn't improve her grades.

a) **nevertheless** b) indeed c) whereas d) as long as e) hence f) given that

Catherine hates hot drinks, _____ she forces herself to drink a cup of herbal tea every night.

a) given that b) hence c) indeed d) **nevertheless** e) as long as f) whereas

Elizabeth likes talking with Christopher very much, _____ she doesn't always pick up when he calls.

a) since b) provided c) therefore d) furthermore e) **nevertheless** f) whereas

Christine is ill, _____ she will go to work anyway.

a) provided b) indeed c) whereas d) therefore e) given that f) **nevertheless**

Despite this (High Frequency)

Christian does not like chocolate, _____ he eats it from time to time.

a) since b) **despite this** c) as long as d) whereas e) indeed f) therefore

Sarah has many colleagues, _____ she knows all their names.

a) provided b) **despite this** c) indeed d) hence e) conversely f) given that

Samuel has always worked a lot, _____ he still doesn't have much money.

a) since b) therefore c) whereas d) **despite this** e) furthermore f) as long as

Anthony always eats a lot, _____ he is very thin.

a) **despite this** b) whereas c) furthermore d) provided e) hence f) since

Lucy failed an exam at school, _____ she is in a good mood.

a) given that b) indeed c) therefore d) whereas e) **despite this** f) as long as

Relation: Contrast

Conversely (Low Frequency)

Florence loves horror films, _____ her husband is disgusted by them.

a) hence b) given that c) **conversely** d) furthermore e) nevertheless f) provided

Evelyn is always late, _____ her brother is always on time.

a) nevertheless b) indeed c) **conversely** d) since e) as long as f) therefore

Victor likes chocolate cakes, _____ his mother prefers fruit cakes.

a) given that b) nevertheless c) hence d) **conversely** e) furthermore f) provided

Alan hates dogs, _____ his brother likes them.

a) therefore b) since c) furthermore d) provided e) nevertheless f) **conversely**

Amadou likes to try out new recipes, _____ his brother never cooks.

a) as long as b) nevertheless c) given that d) hence e) indeed f) **conversely**

Whereas (High Frequency)

Nadir likes his job, _____ his sister would like to change her career.

a) therefore b) since c) **whereas** d) furthermore e) nevertheless f) as long as

Elise really likes to go horseback riding, _____ her brother can't stand animals.

a) indeed b) despite this c) since d) therefore e) provided f) **whereas**

Marion was sad that she wasn't invited to the party, _____ her sister wouldn't have cared either way.

a) **whereas** b) hence c) indeed d) given that e) despite this f) as long as

Marion likes to party a lot, _____ her roommate prefers to stay quiet.

a) as long as b) furthermore c) nevertheless d) **whereas** e) given that f) hence

Alison likes going to the cinema, _____ her friend prefers watching movies on television.

a) provided b) hence c) since d) despite this e) **whereas** f) furthermore

Relation: Condition

Provided (Low Frequency)

Matthew is likely to come, _____ his train arrives on time.

a) **provided** b) nevertheless c) since d) conversely e) indeed f) therefore

Malika would probably go on vacation, _____ her boss agrees to give her two weeks off.

a) hence b) **provided** c) despite this d) given that e) conversely f) indeed

John is going to the beach, _____ the weather holds up.

a) despite this b) given that c) furthermore d) therefore e) **provided** f) conversely

Joanne agreed to read the book, _____ her brother doesn't spoil the ending.

a) furthermore b) whereas c) **provided** d) since e) hence f) despite this

Nicolas will take a few days off, _____ he can convince his wife to do the same.

a) despite this b) given that c) indeed d) **provided** e) whereas f) therefore

As long as (High Frequency)

Pascal can continue living at home, _____ he stays in school.

a) since b) **as long as** c) despite this d) furthermore e) whereas f) therefore

Etienne will be here at 7 o'clock, _____ he doesn't get lost on the way.

a) hence b) conversely c) furthermore d) nevertheless e) given that f) **as long as**

Laura would like to go for a walk tonight, _____ it doesn't start raining.

a) **as long as** b) indeed c) despite this d) since e) conversely f) hence

Laurence will keep his old computer for a few years, _____ it will continue to work properly.

a) indeed b) **as long as** c) despite this d) conversely e) since f) therefore

Anthony will make enough money for the trip, _____ he continues working at his job.

a) despite this b) given that c) conversely d) furthermore e) **as long as** f) therefore

Relation: Cause

Given that (Low Frequency)

Ernest will not bring Claude a gift, _____ he doesn't know him.

a) nevertheless b) conversely c) provided d) indeed e) hence f) **given that**

Jackie must speak Italian very well, _____ she grew up in Italy.

a) furthermore b) **given that** c) therefore d) provided e) whereas f) despite this

Charles couldn't have gone on the hike, _____ he walked with a cane at the time.

a) as long as b) furthermore c) nevertheless d) conversely e) **given that** f) whereas

Sylvia must know the United States well, _____ she is American.

a) nevertheless b) provided c) **given that** d) hence e) conversely f) indeed

I know Chantal will like the concert, _____ she is a musician herself.

a) **given that** b) as long as c) hence d) indeed e) whereas f) conversely

Since (High Frequency)

John stayed home this week, _____ he caught the flu.

a) despite this b) provided c) indeed d) therefore e) **since** f) nevertheless

Alice doesn't like ice cream, _____ she finds it too cold.

a) conversely b) nevertheless c) **since** d) as long as e) therefore f) despite this

Luke wants to become a physicist, _____ he is good at science.

a) conversely b) as long as c) despite this d) **since** e) furthermore f) hence

Jane had to pay a fine, _____ she parked illegally.

a) nevertheless b) **since** c) indeed d) as long as e) conversely f) hence

François must be in very good physical shape, _____ he trains every day.

a) furthermore b) provided c) **since** d) nevertheless e) whereas f) therefore

Table A.2: Connectives by frequency and complexity, with comparisons between the adapted English sentences and answers, as well as the original French version from Wetzel et al. (2020)

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
addition	low	1	Philip has many good qualities—he really likes to cook, _____ he's a good pianist.	**furthermore**	Philipe aime bien cuisiner, _ c'est un bon pianiste	en outre
addition	low	1	Marie and Emily have much in common—they both love gardening, _____ they love sailing.	**furthermore**	Marie est passionnée par le jardinage, _ elle adore la voile.	en outre
addition	low	1	Camilla is a real film fanatic—she likes to go to the cinema, _____ she has a large collection of old movie posters.	**furthermore**	Camille adore aller au stade de foot, _ elle aime voir des vieux films au cinéma.	en outre
addition	low	1	Martin speaks very loudly into the telephone, _____ he never stops talking.	**furthermore**	Martin parle toujours très fort au téléphone, _ il parle sans arrêt.	en outre

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
addition	low	1	Leon seems very cultured—he listens to a lot of classical music, _____ he loves traditional dance.	**furthermore**	Léon écoute beaucoup de musique classique, _ il adore les danses traditionnelles.	en outre
addition	high	1	Gabriel often recites poetry, _____ he was practicing a new poem last night.	**indeed**	Gabriel connaît bien tous les noms d'arbres, _ il est expert en oiseaux.	par ailleurs
addition	high	1	Arthur is always losing his things, _____ he misplaced his keys this morning.	**indeed**	Arthur oublie tout le temps son portable, _ il perd souvent ses clés.	par ailleurs
addition	high	1	Juliet is always ready to help, _____ she volunteers regularly.	**indeed**	Juliette est toujours prête à aider, _ elle est très intelligente.	par ailleurs
addition	high	1	Lucas has made a few enemies, _____ he received a threatening letter recently.	**indeed**	Lucas a beaucoup d'amis, _ il a une grande famille.	par ailleurs
addition	high	1	Emma only drinks black tea, _____ it's her favourite beverage.	**indeed**	Emma boit beaucoup de thé noir, _ elle adore le chocolat	par ailleurs

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
consequence	high	2	Martin avoids doing his mathematics homework, _____ he will struggle on the exam.	**therefore**	Martin évite de manger trop de chocolat, _ il ne prend pas de poids.	ainsi
consequence	high	2	Jane ran fast so as not to miss the train, _____ she was on time for work.	**therefore**	Jeanne courait vite pour ne pas rater son train, _ elle était à l'heure au travail.	ainsi
consequence	high	2	Philip lives next to a grocery store, _____ he has everything he needs to cook.	**therefore**	Philippe habite à côté d'une épicerie, _ il ne manque jamais de rien.	ainsi
consequence	high	2	Nadia often goes running, _____ she's very fit.	**therefore**	Nadia sort souvent courir, _ elle reste en bonne santé.	ainsi
consequence	high	2	Louise works in another city, _____ she takes the train every day.	**therefore**	Louise travaille dans une autre ville, _ elle prend le train tous les jours.	ainsi
consequence	low	2	David was often late when he first started, _____ he always sets his alarm a bit early each morning.	**hence**	Martin est toujours en retard, _ il met son alarme un peu plus tôt chaque matin.	c'est pourquoi

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
consequence	low	2	John doesn't have time to eat a full meal at work, _____ he only takes snacks to the office.	**hence**	Jean n'a souvent pas le temps de manger, _ il emporte des snacks au bureau.	c'est pourquoi
consequence	low	2	Max often forgets his glasses at home, _____ he keeps another pair at work.	**hence**	Max oublie souvent ses lunettes à la maison , _ il en garde une deuxième paire au travail.	c'est pourquoi
consequence	low	2	Jane took private lessons this term, _____ she did well in her exams.	**hence**	Jeanne a pris des cours particuliers ce semestre, _ elle a bien réussi ses examens.	c'est pourquoi
consequence	low	2	Mary is pregnant, _____ she is not drinking alcohol.	**hence**	Marie est enceinte, _ elle a arrêté de travailler.	c'est pourquoi
concession	low	6	Kenneth does not like sports, _____ he does them to stay healthy.	**nevertheless**	Fabien n'aime pas le sport, _ il en fait quand même pour être en bonne santé.	cependant
concession	low	6	Joanne did her homework very well, _____ she couldn't improve her grades.	**nevertheless**	Fabienne a très bien fait ses devoirs , _ elle n'a tout de même pas pu améliorer ses notes.	cependant

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
concession	low	6	Catherine hates hot drinks, _____ she forces herself to drink a cup of herbal tea every night.	**nevertheless**	Catherine déteste les boissons chaudes, _ elle se force quand même à boire une tisane chaque soir.	cependant
concession	low	6	Elizabeth likes talking with Christopher very much, _____ she doesn't always pick up when he calls.	**nevertheless**	Élisabeth aimerait bien parler avec Christophe, _ elle ne décroche tout de même pas quand il appelle.	cependant
concession	low	6	Christine is ill, _____ she will go to work anyway.	**nevertheless**	Christine est malade, _ elle va tout de même aller travailler.	cependant
concession	high	6	Christian does not like chocolate, _____ he eats it from time to time.	**despite this**	Christian n'aime pas le chocolat, _ il en mange quand même de temps en temps.	néanmoins
concession	high	6	Sarah has many colleagues, _____ she knows all their names.	**despite this**	Sarah a de nombreux collègues, _ elle connaît tout de même tous leurs noms.	néanmoins
concession	high	6	Samuel has always worked a lot, _____ he still doesn't have much money.	**despite this**	Samuel a toujours beaucoup travaillé , _ il n'a quand même pas beaucoup d'argent.	néanmoins

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
concession	high	6	Anthony always eats a lot, _____ he is very thin.	**despite this**	Antoine a toujours mangé beaucoup, _ il est tout de même très maigre.	néanmoins
concession	high	6	Lucy failed an exam at school, _____ she is in a good mood.	**despite this**	Lucie a raté un examen à l'école, _ elle est tout de même de bonne humeur.	néanmoins
contrast	low	3	Florence loves horror films, _____ her husband is disgusted by them.	**conversely**	Florence aime les films d'horreur, _ son mari est toujours terrifié.	par contre
contrast	low	3	Evelyn is always late, _____ her brother is always on time.	**conversely**	Fabienne est toujours en retard, _ son frère est toujours à l'heure.	par contre
contrast	low	3	Victor likes chocolate cakes, _____ his mother prefers fruit cakes.	**conversely**	Victor adore les gâteaux au chocolat, _ sa mère préfère les gâteaux aux fruits.	par contre
contrast	low	3	Alan hates dogs, _____ his brother likes them.	**conversely**	Alain déteste les chiens, _ son frère les aime bien.	par contre
contrast	low	3	Amadou likes to try out new recipes, _____ his brother never cooks.	**conversely**	Amadou adore tester des nouvelles recettes, _ son frère ne cuisine jamais.	par contre

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
contrast	high	3	Nadir likes his job, ____ his sister would like to change her career.	**whereas**	Nadir aime son travail, _ sa sœur aimerait changer de carrière.	en revanche
contrast	high	3	Elise really likes to go horseback riding, _____ her brother can't stand animals.	**whereas**	Élise aime beaucoup monter à cheval, _ son frère déteste les animaux.	en revanche
contrast	high	3	Marion was sad that she wasn't invited to the party, _____ her sister wouldn't have cared either way.	**whereas**	Marion n'est pas invitée à la fête, _ sa sœur a reçu une invitation.	en revanche
contrast	high	3	Marion likes to party a lot, _____ her roommate prefers to stay quiet.	**whereas**	Marion aime beaucoup faire la fête, _ sa colocataire préfère rester au calme.	en revanche
contrast	high	3	Alison likes going to the cinema, _____ her friend prefers watching movies on television.	**whereas**	Clément aime aller au cinéma, _ sa copine préfère voir les films à la télévision.	en revanche
condition	low	4	Matthew is likely to come, _____ his train arrives on time.	**provided**	Matthieu va sûrement venir, _ son train soit à l'heure.	pourvu que

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
condition	low	4	Malika would probably go on vacation, _____ her boss agrees to give her two weeks off.	**provided**	Malika partirait en vacances _ sa patronne lui donnait une semaine de congé.	pourvu que
condition	low	4	John is going to the beach, _____ the weather holds up.	**provided**	Jean ira à la plage _ il fasse beau demain.	pourvu que
condition	low	4	Joanne agreed to read the book, _____ her brother doesn't spoil the ending.	**provided**	Jeanne sera d'accord de lire le livre, _ elle puisse le lire avant son frère.	pourvu que
condition	low	4	Nicolas will take a few days off, _____ he can convince his wife to do the same.	**provided**	Nicolas prendra quelques jours de congé, _ il en ait encore cette année.	pourvu que
condition	high	4	Pascal can continue living at home, _____ he stays in school.	**as long as**	Pascal irait faire une promenade, _ il n'avait rien d'autre à faire.	dans le cas où
condition	high	4	Etienne will be here at 7 o'clock, _____ he doesn't get lost on the way.	**as long as**	Etienne viendrait te voir, _ il avait encore du temps après le travail.	dans le cas où

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
condition	high	4	Laura would like to go for a walk tonight, _____ it doesn't start raining.	**as long as**	Laurent irait se promener ce soir, _ il ne pleuvait pas.	dans le cas où
condition	high	4	Laurence will keep his old computer for a few years, _____ it will continue to work properly.	**as long as**	Laurence s'achèterait un nouvel ordinateur _ le sien ne fonctionnait plus.	dans le cas où
condition	high	4	Anthony will make enough money for the trip, _____ he continues working at his job.	**as long as**	Antoine resterait dans son appartement, _ il ne trouvait pas de travail.	dans le cas où
cause	high	5	John stayed home this week, _____ he caught the flu.	**since**	Jean est resté chez lui cette semaine, _ il a attrapé la grippe.	car
cause	high	5	Alice doesn't like ice cream, _____ she finds it too cold.	**since**	Alice n'aime pas les glace, _ elle trouve qu'elles sont trop froides.	car
cause	high	5	Luke wants to become a physicist, _____ he is good at science.	**since**	Luc veut devenir physicien, _ il est très fort en sciences.	car
cause	high	5	Jane had to pay a fine, _____ she parked illegally.	**since**	Jeanne s'est perdue, _ elle ne connaît pas bien la ville.	car

relation	freq	complexity	sentence_en	answer_en	sentence_fr	answer_fr
cause	high	5	François must be in very good physical shape, _____ he trains every day.	**since**	François est en très bonne condition physique, _ il s'entraîne tous les jours.	car
cause	low	5	Ernest will not bring Claude a gift, _____ he doesn't know him.	**given that**	Ernest ne viendra pas visiter Claude, _ il ne le connaît pas.	puisque
cause	low	5	Jackie must speak Italian very well, _____ she grew up in Italy.	**given that**	Jacqueline parle bien l'anglais, _ elle a grandi aux États-Unis.	puisque
cause	low	5	Charles couldn't have gone on the hike, _____ he walked with a cane at the time.	**given that**	Charles ne viendra pas en excursion, _ il a marche avec des cannes en ce moment.	puisque
cause	low	5	Sylvia must know the United States well, _____ she is American.	**given that**	Sylvie connaît bien les États-Unis, _ elle est américaine.	puisque
cause	low	5	I know Chantal will like the concert, _____ she is a musician herself.	**given that**	Chantal va beaucoup aimer le concert, _ elle adore le violoncelle.	puisque

Appendix A.13 Collocations Task

Table A.3: Collocations task stimuli with key ("Words That Go Together" test, Dąbrowska, 2014).

Item	a	b	c	d	e	key
1	**blatant lie**	clear lie	conspicuous lie	distinct lie	recognizable lie	a
2	**blank expression**	frightful expression	plain expression	sinister expression	terrible expression	a
3	attain publicity	**attract publicity**	bring publicity	make publicity	win publicity	b
4	**fair share**	honest share	just share	legitimate share	reasonable share	a
5	**arouse suspicions**	incite suspicions	kindle suspicions	revive suspicions	stimulate suspicions	a
6	elevate prices	grow prices	lift prices	**raise prices**	stimulate prices	d
7	chance a guess	dare a guess	gamble a guess	**hazard a guess**	risk a guess	d
8	**bend rules**	honour rules	institute rules	reject rules	validate rules	a
9	believe a statement	change a statement	**issue a statement**	offer a statement	revise a statement	c
10	advance standards	boost standards	elevate standards	lift standards	**raise standards**	e

Item	a	b	c	d	e	key
11	**boost production**	double production	enlarge production	extend production	redouble production	a
12	combine the ranks	conjoin the ranks	**join the ranks**	merge the ranks	unify the ranks	c
13	**bitter dispute**	cruel dispute	hard dispute	harsh dispute	savage dispute	a
14	**absolute silence**	pure silence	sheer silence	stark silence	supreme silence	a
15	complete confession	exhaustive confession	extensive confession	**full confession**	thorough confession	d
16	acquire popularity	attract popularity	earn popularity	**gain popularity**	get popularity	d
17	constant employment	normal employment	ordinary employment	**regular employment**	unbroken employment	d
18	glimpse an incident	notice an incident	observe an incident	see an incident	**witness an incident**	e
19	**achieve one's objectives**	complete one's objectives	finish one's objectives	follow one's objectives	tackle one's objectives	a
20	accurate direction	appropriate direction	convenient direction	**general direction**	specific direction	d
21	apply attention	dedicate attention	**divert attention**	grasp attention	sidetrack attention	c
22	extensive problem	extreme problem	**serious problem**	significant problem	vital problem	c

Item	a	b	c	d	e	key
23	compelling matters	critical matters	desperate matters	major matters	**urgent matters**	e
24	**close similarity**	doubtful similarity	evident similarity	extreme similarity	near similarity	a
25	contradict rumours	discover rumours	**hear rumours**	know rumours	tell rumours	c
26	effective phrase	helpful phrase	**memorable phrase**	noteworthy phrase	significant phrase	c
27	distract suspicion	**divert suspicion**	mislead suspicion	redirect suspicion	sidetrack suspicion	b
28	bring faith	instil faith	offer faith	refresh faith	**restore faith**	e
29	complete search	full search	scrupulous search	**thorough search**	total search	d
30	abundant details	complete details	definite details	**precise details**	small details	d
31	apply punishment	deliver punishment	**inflict punishment**	perform punishment	provide punishment	c
32	appealing proposition	**attractive proposition**	charming proposition	inviting proposition	seductive proposition	b
33	dark view	**dim view**	murky view	shadowy view	shady view	b
34	aggressive critic	forthright critic	frank critic	open critic	**outspoken critic**	e
35	**odd remark**	peculiar remark	queer remark	unnatural remark	weird remark	a

Item	a	b	c	d	e	key
36	distinct example	gross example	recognizable example	shocking example	**striking example**	e
37	formulate a complaint	**lodge a complaint**	place a complaint	record a complaint	write a complaint	b
38	confident conclusion	evident conclusion	**obvious conclusion**	solid conclusion	sure conclusion	c
39	general responsibility	large responsibility	**overall responsibility**	single responsibility	unique responsibility	c
40	decline an application	deny an application	ignore an application	**refuse an application**	scrap an application	d

Appendix A.14 Idiomatic phrases and CQL search terms

Table A.4: Idiomatic phrases and CQL search terms.

Legend: "fam." = "familiarity", "plaus." = literal plausibility; "BNC SP ppm" = BNC Spoken Corpus parts-per-million; "BNC WR ppm" = BNC Written Corpus parts-per-million. BNC values calculated from the BNC corpus in SketchEngine (Kilgarriff et al., 2014). Familiarity and literal plausibility values, where available, are from a database of English idiom norms (Bulkes & Tanner, 2017).

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
(open) **a can of worms**	0.31	0.42	0.3	3.03	4.03	phrase: can of worms
(to go on) **a wild goose chase**	0.16	NA	0.18	3.49	4.20	[word="wild"] [word="goose"] [lemma="chase"]
the **bloom is off the rose**	0.01	NA	0.01	NA	NA	[word="bloom"] [word="was"] []{0,6} [word="off"] [word="the"] [word="rose"]
to **be (left) out of pocket**	0.58	0.51	0.59	NA	NA	[word="out"] [word="of"] [word="pocket"]
to **be cannon fodder**	0.14	0.08	0.15	1.96	1.96	[word="cannon"] [word="fodder"]
to **be given the boot**	0.15	0.08	0.16	3.22	4.27	[lemma="give" & tag="V.*"] []{0,3} [word="the"] [word="boot"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **be no spring chicken**	0.05	NA	0.06	3.00	2.16	phrase: no spring chicken
to **be on the edge of one's seat**	0.22	0.17	0.23	3.96	4.81	[lemma="be" & tag="V.*"] [word="on"] [word="the"] [word="edge"] [word="of"] [tag="PP.?"] [word="seat"]
to **be over the moon**	0.5	0.59	0.49	2.65	1.80	[lemma="be" & tag="V.*"] [lemma="over"] [lemma="the"] [lemma="moon"]
to **be sawing logs**	NA	NA	NA	NA	NA	[word="sawing"] [lemma="log"]
to **be skating on thin ice**	0.07	0.17	0.06	3.53	4.71	[lemma="skate" & tag="V.*"] [word="on"] []{0,3} [word="thin"] [word="ice"]
to **be under the weather**	0.28	0.17	0.29	4.20	1.75	phrase: under the weather
to **be used as a guinea pig**	0.24	0.25	0.24	3.25	1.96	([lemma="be" & tag="VB.*"] | [lemma="make" & tag="VB.*"]) []{0,3} [word="used"]? [word="as"]? [word="as"]? [word="a"]? [lemma="guinea" & tag="NN.*"] [lemma="pig" & tag="NN.*"]
to **beat someone off with a stick**	NA	NA	NA	NA	NA	[lemma="beat" & tag="V.*"] [tag="PP.?"]? [word= "off"] []{0,3} [word= "with"] [word= "a"] [word= "stick"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **bite off more than one can chew**	0.08	NA	0.09	NA	NA	[lemma="bite" & tag="V.*"] [lemma="off"] [lemma="more"] [lemma="than"] [tag="PP.?"] [lemma="can"] [lemma="chew"]
to **break (through) the glass ceiling**	0.02	NA	0.02	4.08	1.23	[lemma="break" & tag="V.*"] [word="through"]? [word="the"] [word="glass"] [word="ceiling"]
to **brush up (on something)**	0.33	0.93	0.26	NA	NA	[lemma="brush" & tag="V.*"] [word="up"] ([tag="PP.?"] |[word="on"])
to **burn the candle at both ends**	0.05	0.08	0.05	3.13	3.43	[lemma="burn" & tag="V.*"] [word="the"] []{0,2} [lemma="candle"] [word="at"] [word="both"] [word="ends"]
to **burn the midnight oil**	0.09	NA	0.1	2.63	3.65	[lemma="burn" & tag="V.*"] [word="the"] [lemma="midnight"] [lemma="oil"]
to **bury one's head in the sand**	0.09	0.17	0.08	3.17	3.92	[lemma="bury" & tag="V.*"] [tag="PP.?"] [word= "head"] [word= "in"] [word= "the"] [word= "sand"]
to **bury the hatchet**	0.12	NA	0.14	3.52	4.37	[lemma="bury" & tag="V.*"] []{0,3} [lemma = "hatchet"]
to **carry a torch for someone**	0.03	NA	0.03	2.63	4.21	[lemma="carry" & tag="V.*"] [word="a"] [word="torch"] [word="for"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **come out of one's shell**	0.07	NA	0.08	3.69	2.55	[lemma="come" & tag="V.*"] [word="out"] [word="of"] [tag="PP.?"] [word="shell"]
to **come to a crossroads**	0.04	NA	0.05	3.19	4.72	[lemma="come" & tag="V.*"] [word="to"] [word="a"] []{0,3} [word="crossroads"]
to **dig one's own grave**	0.05	NA	0.06	3.57	4.37	[lemma="dig" & tag="V.*"] [tag="PP.?"] [word="own"] [word="grave"]
to **drive someone up the wall**	0.11	0.68	0.04	3.88	1.93	[lemma="drive" & tag="V.*"] [tag="PP.?"] [word="up"] [word="the"] []{0,3} [word="wall"]
to **drop a/the bombshell**	0.1	NA	0.11	3.35	3.47	[lemma="drop" & tag="V.*"] [word="the"] [lemma="bombshell"]
to **face the music**	0.28	NA	0.32	3.29	3.18	[lemma="face" & tag="V.*"] [word="the"] [word="music"]
to **fall off the wagon**	NA	NA	NA	3.43	4.43	[lemma="fall" & tag="V.*"] [word = "off"] [word = "the"] []{0,3} [word = "wagon"]
to **fly the coop**	0.01	NA	0.01	2.78	2.35	[lemma="fly" & tag="V.*"] [word="the"] [word="coop"]
to **get wind of something**	0.38	0.17	0.41	3.55	1.79	[lemma="get" & tag="V.*"] [word="wind"] [word="of"]
to **go bananas**	0.2	0.59	0.15	3.25	1.14	[lemma="go" & tag="V.*"] [word="bananas"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **go off the deep end**	0.02	NA	0.02	3.88	3.84	[lemma="go" & tag="V.*"] [word="off"] [word="the"] [word="deep"] [word="end"]
to **go through something with a fine tooth comb**	0.06	0.34	0.03	3.39	4.08	[lemma="go" & tag="V.*"] [word="through"] [tag="PP.?"] [word="with"] [word="a"] [word="fine"] [word="toothcomb|tooth"] [word="comb"]?
to **go under the knife**	0.02	NA	0.02	3.32	4.08	[lemma="go" & tag="V.*"] [word="under"] [word="the"] [word="knife"]
to **grease someone's palms**	0.01	NA	0.01	NA	NA	[lemma="grease" & tag="V.*"] [tag="PP.?"] [word="palms"]
to **have a bone to pick with someone**	0.02	NA	0.02	NA	NA	[lemma="have" & tag="V.*"] [word="a"] [word="bone"] [word="to"] [word="pick"]
to **have a few screws loose**	0.12	0.42	0.09	3.83	2.87	([tag="N.*"]|[tag="PP.?"]) []{0,4} [lemma="screw" & tag="N.*"] [word="loose"]
to **have a sweet tooth**	0.24	0.08	0.26	4.30	1.80	[word="a"] [word="sweet"] [word="tooth"]
to **hit the hay**	0.01	NA	0.01	3.16	4.08	[lemma="hit" & tag="V.*"] []{0,3} [lemma = "hay"]
(much/ a lot of) **ink has been spilled over something**	0.04	NA	0.04	NA	NA	[word="ink"] [lemma="have"] []{0,1} [lemma="spill" & tag="V.*"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **kick the bucket**	0.11	0.42	0.07	NA	NA	[lemma="kick" & tag="V.*"] [word="the"] [word="bucket"]
to **lend an ear (to someone)**	0.1	NA	0.11	2.98	1.49	[lemma="lend" & tag="V.*"] [tag="PP.?"]? []{0,3} [word="ear"]
to **let the cat out of the bag**	0.12	0.08	0.12	NA	NA	[lemma="let" & tag="V.*"] [word="the"] [lemma = "cat"] [lemma = "out"] [lemma = "of"] [lemma = "the"] [lemma = "bag"]
to **lose one's thread**	0.07	0.25	0.05	NA	NA	[lemma="lose" & tag="V.*"] [tag="PP.?"] [lemma = "thread"]
to **make a killing**	0.28	0.25	0.28	4.03	2.75	[lemma="make" & tag="V.*"] [word="a"] [word="killing"]
to **make a pass at someone**	0.2	0.08	0.22	NA	NA	[lemma="make" & tag="V.*"] [word="a"]? [word="pass"] [word="at"] [tag="PP.?"]
to **not be someone's cup of tea**	0.08	0.25	0.06	3.82	4.18	[word="not"] [tag="PP.?"] [word="cup"] [word="of"] [word="tea"]
to **pick someone's brain**	0.02	NA	0.02	3.83	1.91	[lemma="pick" & tag="V.*"] [tag="PP.?"] [word="brain"]
to **play chicken**	0.02	NA	0.02	2.92	2.58	[lemma="play" & tag="V.*"] [word="chicken"]
to **play the field**	0.13	NA	0.15	3.48	3.19	[lemma="play" & tag="V.*"] [word="the"] [word="field"]
to **push someone's buttons**	0.01	NA	0.01	4.03	2.95	[lemma="push" & tag="V.*"] [tag="PP.?"] [lemma="button"]

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **push the boundaries**	0.23	0.25	0.23	NA	NA	[lemma="push" & tag="V.*"] []{0,3} [lemma="boundary"]
to **put down roots**	0.16	NA	0.18	2.56	2.80	[lemma="put" & tag="V.*"] [word= "down"] [word= "roots"]
to **put oneself in someone else's shoes**	0.06	0.08	0.06	4.03	4.07	[lemma="put" & tag="V.*"] [tag="PP.?"] [word="in"] [tag="PP.?"] [word="else's"]? [word="shoes"]
to **quake in one's boots**	0.01	0.08	NA	2.61	3.53	[lemma="quake" & tag="V.*"] [word="in"] [tag="PP.?"] ([lemma="boot"] | [lemma="shoe"])
to **rake/haul someone over the coals**	0.1	NA	0.11	2.74	2.83	[tag="V.*"] [tag="PP.?"]? [word="over"] [word="the"] [word="coals"]
to **read too much into something**	0.28	0.25	0.29	4.08	2.65	[lemma="read" & tag="V.*"] [word="too"] [word="much"] [word="into"]
to **roll out/expect/give/have the red carpet**	0.13	NA	0.15	3.42	4.70	([lemma="expect" & tag="V.*"] | [lemma="roll" & tag="V.*"] | [lemma="give" & tag="V.*"] | [lemma="have" & tag="V.*"] | [lemma="lay" & tag="V.*"]) [] {0,3} [lemma="the"] [lemma="red"] [lemma="carpet"]
to **take a rain check**	0.04	0.08	0.04	3.78	2.76	[word="raincheck"] | ([word="rain"] [word="check"])

phrase	BNC ppm	BNC SP ppm	BNC WR ppm	fam.	plaus.	CQL/phrase search
to **thank/count one's lucky stars**	0.18	NA	0.2	3.58	2.29	[word="lucky"] [word="stars"]
to **tie the knot**	0.34	NA	0.38	NA	NA	[lemma="tie" & tag="V.*"] [word="the"] [word="knot"]
to **walk on eggshells**	0.02	NA	0.02	3.73	3.99	[lemma="walk" & tag="V.*"] [word="on"] [word="eggshells"]

Appendix A.15 Visual world paradigm

Table A.5: Visual world paradigm sentences by version and trial type (idi = idiomatic; lit = literal), with file names for image options.
Note: "A/B" sentences are practice trials. The initial bolded section of each sentence indicates the onset of the figurative or literal phrase, the second bolded section indicates onset of the critical word or phrase.

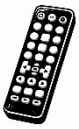
ver.	type	sentence	target	foil	dist_1	dist_2
A/B	idi	He was **over the moon,** and I can honestly say I have never before seen him look so **happy.**	 happy.jpg	 moon.jpg	 phone.jpg	 backpack.jpg
A/B	lit	He was **under the moon,** and he was looking with amazement through his **powerful telescope.**	 moon.jpg	 happy.jpg	 phone.jpg	 backpack.jpg
A/B	idi	She **read too much into the comment,** and it made her feel extremely **angry.**	 anger_woman_2.jpg	 reading.jpg	 ufo.jpg	 drum.jpg
A/B	lit	She **read too much into the night,** and she wound up finishing the entire **book.**	 reading.jpg	 anger_woman_2.jpg	 ufo.jpg	 drum.jpg

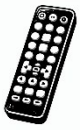
ver.	type	sentence	target	foil	dist_1	dist_2
A/B	idi	The **bloom was off the rose,** and I wondered if we were soon going to **break up.**	 disinterest.jpg	 rose.jpg	 basket.jpg	 headphones.jpg
A/B	lit	The **petals fell off the rose,** and I wondered if we were soon going to buy a new **bouquet.**	 rose.jpg	 disinterest.jpg	 basket.jpg	 headphones.jpg
A	idi	He **opened a can of worms,** and it wound up creating an even bigger **problem.**	 facepalm_2.jpg	 worms.jpg	 circus.jpg	 fingerprint.jpg
B	lit	He opened a **box of worms,** which he intended to use for **fishing.**	 worms.jpg	 facepalm_2.jpg	 circus.jpg	 fingerprint.jpg
B	idi	It was time to **bury the hatchet,** and soon they became the **closest of friends.**	 friends_2.jpg	 hatchet.jpg	 mushrooms.jpg	 jacket.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	It was time to **store the hatchet,** so he put it out of reach in the **shed.**	 hatchet.jpg	 friends_2.jpg	 mushrooms.jpg	 jacket.jpg
A	idi	She decided to **hit the hay,** and in a few moments she was **asleep.**	 asleep_woman.jpg	 haystack_2.jpg	 fan.jpg	 shark.jpg
B	lit	She decided to **move the hay,** and placed it on the other side of the **barn.**	 haystack_2.jpg	 asleep_woman.jpg	 fan.jpg	 shark.jpg
B	idi	She **let the cat out of the bag,** and now everyone knows the awful **secret.**	 secret_2.jpg	 cat.jpg	 snail.jpg	 beer.jpg
A	lit	She **let the cat out of the house,** and it never came back **home.**	 cat.jpg	 secret_2.jpg	 snail.jpg	 beer.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	He **lost his thread,** and it was clear to everyone that he was extremely **confused.**	 confused.jpg	 thread.jpg	 goat.jpg	 kite.jpg
B	lit	She **pulled the thread,** and before long, she had ruined her favourite **sweater.**	 thread.jpg	 confused.jpg	 goat.jpg	 kite.jpg
B	idi	They decided to **tie the knot,** and the next weekend they were **married.**	 marriage.jpg	 knot2.jpg	 butterfly.jpg	 camera.jpg
A	lit	They decided to **secure the knot,** in order to support their combined **weight.**	 knot2.jpg	 marriage.jpg	 butterfly.jpg	 camera.jpg
A	idi	He finally **kicked the bucket,** but I couldn't believe he was really **dead.**	 grave.jpg	 bucket_3.jpg	 luggage.jpg	 ant.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	He finally **filled the bucket,** but before doing so, he lost a lot of **milk.**	 bucket_3.jpg	 grave.jpg	 luggage.jpg	 ant.jpg
B	idi	He was **given the boot,** but he had a hard time telling his family he was **fired.**	 fired.jpg	 boot.jpg	 tree.jpg	 #starfish.jpg
A	lit	He was **fixing the boot,** but he had a hard time putting on the **rubber sole.**	 boot.jpg	 fired.jpg	 tree.jpg	 starfish.jpg
A	idi	He would **go bananas** whenever he was feeling **upset.**	 anger_man_3.jpg	 bananas.jpg	 horse.jpg	 wateringcan.jpg
B	lit	He would **choose bananas** whenever he was feeling **hungry.**	 bananas.jpg	 anger_man_3.jpg	 horse.jpg	 wateringcan.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	Last night I **went on a wild goose chase,** trying unsuccessfully to track down the **missing files.**	 searching_3 .jpg	 goose.jpg	 guitar.jpg	 hotairballoon .jpg
A	lit	Last week I **went on a wild goose hunt,** trying unsuccessfully to track the **birds.**	 goose.jpg	 searching_3 .jpg	 guitar.jpg	 hotairballoon .jpg
A	idi	She **had a sweet tooth,** so she always seemed to have room for **dessert.**	 cake.jpg	 tooth_2.jpg	 necktie.jpg	 laptop.jpg
B	lit	She **had a sore tooth,** so she had to find a way to carefully **brush.**	 tooth_2.jpg	 cake.jpg	 necktie.jpg	 laptop.jpg
B	idi	She wanted to **pick his brain,** so she asked when they could have a **conversation.**	 conversation .jpg	 brain.jpg	 remote.jpg	 carrot.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	She wanted to **examine his brain,** since she was worried about his **symptoms.**	 brain.jpg	 conversation .jpg	 remote.jpg	 carrot.jpg
A	idi	This **wasn't his cup of tea,** and he clearly showed his **disapproval.**	 disapproval.jpg	 mug.jpg	 controller.jpg	 train.jpg
B	lit	This **wasn't his cup of coffee,** and he wondered if his had already been **washed.**	 mug.jpg	 disapproval.jpg	 controller.jpg	 train.jpg
B	idi	They were caught **playing chicken,** and they were lucky no one was hurt in a **crash.**	 racing.jpg	 chicken.jpg	 hammer.jpg	 web.jpg
A	lit	They were caught **eating chicken,** and they had to admit they weren't really **vegetarians.**	 chicken.jpg	 racing.jpg	 hammer.jpg	 web.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	She knew how to **push his buttons,** causing him to become increasingly **annoyed.**	 annoyed.jpg	 button.jpg	 baby.jpg	 bridge.jpg
A	lit	She knew how to press the buttons in order to activate the code.	 button.jpg	 annoyed.jpg	 baby.jpg	 bridge.jpg
B	idi	She **raked him over the coals,** delivering some harsh but well-deserved **criticism.**	 scold_3.jpg	 barbecue.jpg	 socks.jpg	 bread.jpg
A	lit	She **roasted it over the coals,** which made for the most delicious **barbecue.**	 barbecue.jpg	 scold_3.jpg	 socks.jpg	 bread.jpg
A	idi	They **rolled out the red carpet,** impressing the visitors with the expensive **gala.**	 gala.jpg	 carpet.jpg	 octopus.jpg	 flashlight.jpg

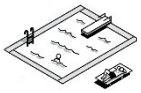
ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	They **rolled out the brown carpet,** trying to determine if it had been **cleaned.**	 carpet.jpg	 gala.jpg	 octopus.jpg	 flashlight.jpg
B	idi	He **went through it with a fine-toothed comb,** but there were never any **clues.**	 searching.jpg	 comb.jpg	 fireplace.jpg	 pig.jpg
A	lit	He **went through it with a nice thin comb,** but he couldn't **untangle her hair.**	 comb.jpg	 searching.jpg	 fireplace.jpg	 pig.jpg
A	idi	She was **on the edge of her seat,** watching the events unfold with increasing **excitement.**	 excited.jpg	 seat.jpg	 leaf.jpg	 handshake.jpg
B	lit	She was **on the front of her seat,** waiting for the next **bus stop.**	 seat.jpg	 excited.jpg	 leaf.jpg	 handshake.jpg

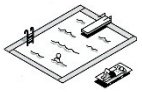
ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	He **got wind of it,** but we're not sure how he could have heard the **news.**	 finding_out_2.jpg	 wind.jpg	 thief.jpg	 dog.jpg
A	lit	He **felt the wind of the ocean,** and he didn't at all mind the **strong breeze.**	 wind.jpg	 finding_out_2.jpg	 thief.jpg	 dog.jpg
A	idi	He **made a killing,** and in just a few months' time he had doubled his **investment.**	 money.jpg	 murder.jpg	 tv.jpg	 mouth.jpg
B	lit	He **arranged the killing,** and in just a few months' time he had been put in **prison.**	 murder.jpg	 money.jpg	 tv.jpg	 mouth.jpg
B	idi	He tried to **put himself in her shoes,** hoping to understand her **feelings.**	 thinking.jpg	 shoes.jpg	 bus.jpg	 treasure.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	He tried to **put on her shoes,** before realising they were the wrong **pair.**	 shoes.jpg	 thinking.jpg	 bus.jpg	 treasure.jpg
A	idi	He **thanked his lucky stars,** feeling a newfound sense of **relief.**	 relief_2.jpg	 stars.jpg	 snowman.jpg	 rice.jpg
B	lit	He **spotted the brightest stars,** trying out his brand new **telescope.**	 stars.jpg	 relief_2.jpg	 snowman.jpg	 rice.jpg
B	idi	He **walked on eggshells around her,** and you could increasingly sense his **nervousness.**	 tense.jpg	 eggshells.jpg	 mountain.jpg	 vulture.jpg
A	lit	He **gathered the eggshells for her,** and put them into the **compost.**	 eggshells.jpg	 tense.jpg	 mountain.jpg	 vulture.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	He **drove her up the wall,** and everyone could see she found him **annoying.**	 annoyed_2.jpg	 car.jpg	 tower.jpg	 pinwheel.jpg
B	lit	He **drove her down the street,** and when they arrived, she hopped out of the **car.**	 car.jpg	 annoyed_2.jpg	 tower.jpg	 pinwheel.jpg
B	idi	He had **come to a crossroads,** and he realised that he faced a very difficult **decision.**	 decision.jpg	 intersection.jpg	 rhino.jpg	 stink.jpg
A	lit	He had **driven to a crossroads,** and he realised that he had forgotten where to **turn.**	 intersection.jpg	 decision.jpg	 rhino.jpg	 stink.jpg
A	idi	She **came out of her shell,** and by the end of the summer she had made a new **friend.**	 friends.jpg	 shell.jpg	 crow.jpg	 skateboard.jpg

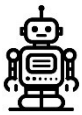
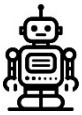
ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	She had **lost her shell,** her only souvenir of her wonderful **summer at the beach.**	 shell.jpg	 friends.jpg	 crow.jpg	 skateboard.jpg
B	idi	She **carried a torch for him,** even though he was unaware of her **romantic feelings.**	 love.jpg	 torch.jpg	 bellpepper.jpg	 elephant.jpg
A	lit	She **brought a torch for him,** knowing that he would need to have the **light.**	 torch.jpg	 love.jpg	 bellpepper.jpg	 elephant.jpg
A	idi	He **dug his own grave,** so he blames himself for his **humiliation.**	 shame3.jpg	 grave_2.jpg	 wrench.jpg	 frog_2.jpg
B	lit	He **picked his own grave,** so he knew he was soon going to **die.**	 grave_2.jpg	 shame3.jpg	 wrench.jpg	 frog_2.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	She **went off the deep end,** and we were all a little surprised by her **explosive anger.**	 anger_x.jpg	 diving.jpg	 artwork.jpg	 igloo.jpg
A	lit	She **jumped off the deep end,** and we were all a little surprised by her **enormous splash.**	 diving.jpg	 anger_x.jpg	 artwork.jpg	 igloo.jpg
A	idi	He has to **grease their palms,** since they always seem to expect a **bribe.**	 money_3.jpg	 hands.jpg	 stump.jpg	 aeroplane.jpg
B	lit	He has to **moisturise his palms,** since he is constantly washing his **hands.**	 hands.jpg	 money_3.jpg	 stump.jpg	 aeroplane.jpg
B	idi	She **had a bone to pick with him,** so she let him know why she was so **upset.**	 anger_woman.jpg	 bone.jpg	 sailboat.jpg	 pool.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	She **had a bone to give to him,** in hopes he could include it in the **museum.**	 bone.jpg	 anger_woman .jpg	 sailboat.jpg	 pool.jpg
A	idi	They tried to **push the boundaries** even further with the latest **fashions.**	 fashion.jpg	 map.jpg	 gorilla.jpg	 fingers_ crossed.jpg
B	lit	They tried to **mark the boundaries** more clearly on the latest **maps.**	 map.jpg	 fashion.jpg	 gorilla.jpg	 fingers_ crossed.jpg
B	idi	He had to **go under the knife,** and it took him months to recover from the **surgery.**	 surgery.jpg	 knife.jpg	 icecream.jpg	 rainbow.jpg
A	lit	He had to **pick up the knife** if he hoped to complete the detailed **etching.**	 knife.jpg	 surgery.jpg	 icecream.jpg	 rainbow.jpg

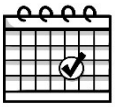
ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	He was **sawing logs,** and it would be many hours before he would **wake up.**	 asleep_4.jpg	 log.jpg	 meeting.jpg	 mask.jpg
B	lit	He was **chopping logs,** trying to gather as much as possible to use as **firewood.**	 log.jpg	 asleep_4.jpg	 meeting.jpg	 mask.jpg
B	idi	He knew he had to **face the music,** so with some hesitation, he decided to **surrender.**	 surrender.jpg	 music.jpg	 eyes.jpg	 ladder.jpg
A	lit	She knew she had to **play the music,** so with some hesitation, she decided to **practice.**	 music.jpg	 surrender.jpg	 eyes.jpg	 ladder.jpg
A	idi	He **dropped the bombshell,** and it left everyone utterly **shocked.**	 shock.jpg	 bombshell.jpg	 keys.jpg	 corn.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	He **launched the bombshell,** and it left the village completely **destroyed.**	 bombshell.jpg	 shock.jpg	 keys.jpg	 corn.jpg
B	idi	She must be **burning the midnight oil,** lately she looks so **tired.**	 tired_3.jpg	 lantern.jpg	 bunny.jpg	 scissors.jpg
A	lit	She must be **burning the lantern oil,** in fact I can already see its **light.**	 lantern.jpg	 tired_3.jpg	 bunny.jpg	 scissors.jpg
A	idi	She was **burning the candle at both ends,** and you could tell she was **exhausted.**	 tired.jpg	 candles.jpg	 shrug.jpg	 jeans.jpg
B	lit	She was **selling the candle at both shops,** but at the other one, they were **bigger.**	 candles.jpg	 tired.jpg	 shrug.jpg	 jeans.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	She was **no spring chicken,** but she couldn't admit she was getting **older.**	 elderly.jpg	 chicken_2.jpg	 halloween.jpg	 saxophone.jpg
B	lit	She was **no ordinary chicken,** she was twice as big as the other **hens.**	 chicken_2.jpg	 elderly.jpg	 halloween.jpg	 saxophone.jpg
A	idi	He was **under the weather,** so he decided to report that he was **ill.**	 illness.jpg	 sun.jpg	 robot.jpg	 ticket.jpg
B	lit	He was **loving the weather,** since he wasn't bothered at all by the **sun.**	 sun.jpg	 illness.jpg	 robot.jpg	 ticket.jpg
B	idi	A lot of **ink has been spilled over this,** even from some of the more reputable **newspapers.**	 newspaper.jpg	 ink.jpg	 hat.jpg	 puppet.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	A lot of **paint has been spilled over here,** and I think it's coming from your **paintbrush.**	 ink.jpg	 newspaper.jpg	 hat.jpg	 puppet.jpg
A	idi	She **broke through the glass ceiling,** after a few years of fighting for a **promotion.**	 promotion.jpg	 glass.jpg	 embarassed.jpg	 tractor.jpg
B	lit	She **broke through the glass window,** after a few seconds of using the **hammer.**	 glass.jpg	 promotion.jpg	 embarassed.jpg	 tractor.jpg
B	idi	He **made a pass at her,** and judging by the response, she didn't appreciate his **flirtation.**	 flirt.jpg	 football.jpg	 bear.jpg	 globe.jpg
A	lit	He **caught a pass from her,** and within a few seconds, **he was able to score the goal.**	 football.jpg	 flirt.jpg	 bear.jpg	 globe.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	She was **left out of pocket,** and she wasn't sure when she was going to have the **money.**	 money_2.jpg	 pocket.jpg	 canoe.jpg	 jellyfish.jpg
B	lit	She was **looking in her pocket,** hoping desperately that she still had her **passport.**	 pocket.jpg	 money_2.jpg	 canoe.jpg	 jellyfish.jpg
B	idi	He had **flown the coop,** but they couldn't figure out how he had **escaped.**	 runningaway.jpg	 coop.jpg	 turtle.jpg	 trumpet.jpg
A	lit	He had **built the coop,** but he couldn't figure out how to **raise chickens.**	 coop.jpg	 runningaway.jpg	 turtle.jpg	 trumpet.jpg
A	idi	He always **played the field,** not wanting to develop a long-term **relationship.**	 dating.jpg	 field.jpg	 submarine.jpg	 rocket.jpg

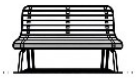
ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	He always **ploughed the field,** ensuring there would be enough for a **bountiful harvest.**	 field.jpg	 dating.jpg	 submarine.jpg	 rocket.jpg
B	idi	He **lent an ear to her,** knowing that she first needed to be **consoled.**	 consolation.jpg	 ear.jpg	 apple.jpg	 book.jpg
A	lit	She **tilted an ear to him,** and it was then that he realised that she was slightly **deaf.**	 ear.jpg	 consolation.jpg	 apple.jpg	 book.jpg
A	idi	He **took a rain check,** asking if it would be possible to **meet on another day.**	 calendar.jpg	 rain.jpg	 astronaut.jpg	 guitar.jpg
B	lit	He **took a rain coat,** anticipating that before the end of the day, there would be a **storm.**	 rain.jpg	 calendar.jpg	 astronaut.jpg	 guitar.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	He **fell off the wagon,** which led to a long night of excessive **drinking.**	 alcohol.jpg	 wagon.jpg	 lightbulb.jpg	 printer.jpg
A	lit	He **jumped off the wagon,** which led to a long night in the nearby **hospital.**	 wagon.jpg	 alcohol.jpg	 lightbulb.jpg	 printer.jpg
A	idi	He was **quaking in his boots,** and it occurred to me that he may be genuinely **terrified.**	 worried.jpg	 boots_2.jpg	 fridge.jpg	 giraffe.jpg
B	lit	He was **splashing in his boots,** with no regard whatsoever about getting **wet feet.**	 boots_2.jpg	 worried.jpg	 fridge.jpg	 giraffe.jpg
B	idi	She really **needs to brush up,** I'm concerned she may not do well on her **exam.**	 study.jpg	 toothbrush.jpg	 windmill.jpg	 crown.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	lit	She really **needs to brush,** I'm concerned she may have a number of **cavities.**	 toothbrush.jpg	 study.jpg	 windmill.jpg	 crown.jpg
A	idi	They were **cannon fodder,** but they accepted this possibility when they decided to be **soldiers.**	 soldiers_2.jpg	 cannon.jpg	 easy.jpg	 snake.jpg
B	lit	They were **cannon projectiles,** and they were typically made out of solid **iron.**	 cannon.jpg	 soldiers_2.jpg	 easy.jpg	 snake.jpg
B	idi	I had to **beat them off with a stick,** I've never seen such an **enthusiastic crowd.**	 crowd.jpg	 stick.jpg	 sad.jpg	 can.jpg
A	lit	I had to **knock them off with a stick,** it was the only way to safely remove the **wasp nests.**	 stick.jpg	 crowd.jpg	 sad.jpg	 can.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
A	idi	He **buried his head in the sand,** but it wouldn't go away simply by **ignoring it.**	 ignoring.jpg	 sand.jpg	 zebra.jpg	 piggybank.jpg
B	lit	He **buried his gun in the sand,** thinking no one would ever dig **it up.**	 sand.jpg	 ignoring.jpg	 zebra.jpg	 piggybank.jpg
B	idi	She **bit off more than she could chew,** and she found herself completely **overwhelmed**.	 overwhelmed .jpg	 bite.jpg	 yoga.jpg	 pineapple.jpg
A	lit	She **bit off more to chew,** and she remained silent for the rest of the **dinner.**	 bite.jpg	 overwhelmed .jpg	 yoga.jpg	 pineapple.jpg
A	idi	She decided to **put down roots,** and within a few months she had found a **house.**	 house_sold.jpg	 roots.jpg	 dragon.jpg	 viking.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	lit	She decided to **pull out the roots,** even though she had never before tried **gardening.**	 roots.jpg	 house_sold.jpg	 dragon.jpg	 viking.jpg
B	idi	He was **used as a guinea pig,** and so he would always try the latest **treatment.**	 syringe.jpg	 guineapig.jpg	 flirt_5.jpg	 notepad.jpg
A	lit	He was **given a guinea pig,** since he always loved playing with **small animals.**	 guineapig.jpg	 syringe.jpg	 flirt_5.jpg	 notepad.jpg
A	idi	**She was skating on thin ice,** and she knew that if she wasn't careful, she could be **fired.**	 anxious_at_work.jpg	 skating.jpg	 thankyou2.jpg	 burger.jpg
B	lit	She was **skating on slippery ice,** and she knew that if she wasn't careful, she could **fall.**	 skating.jpg	 anxious_at_work.jpg	 thankyou2.jpg	 burger.jpg

ver.	type	sentence	target	foil	dist_1	dist_2
B	idi	I knew he **had a few screws loose,** but I didn't think he was completely **mad**.	 mad.jpg	 screws.jpg	 bench.jpg	 mouse.jpg
A	lit	I knew he had **misplaced a few screws,** but I thought he had enough to repair the **shelf.**	 screws.jpg	 mad.jpg	 bench.jpg	 mouse.jpg

Appendix B: Results

Appendix B.1 Author Recognition Tests

Appendix B.1.1 English ARTs – L1 English Samples

Table B.1: Chapter 2, English ART author selections by L1 English University of Oxford student participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	J.K. Rowling	652.09	49	98.00
2	Roald Dahl	692.75	47	94.00
3	Stephen King	698.68	46	92.00
4	George Orwell	693.55	46	92.00
5	Agatha Christie	756.09	46	92.00
6	Virginia Woolf	808.35	44	88.00
7	J.R.R. Tolkien	705.15	41	82.00
8	John Green	776.93	40	80.00
9	T.S. Eliot	830.96	40	80.00
10	Ernest Hemingway	847.53	39	78.00
11	F. Scott Fitzgerald	895.93	38	76.00
12	Harper Lee	822.69	38	76.00
13	Jacqueline Wilson	790.72	37	74.00
14	David Walliams	861.41	37	74.00
15	Margaret Atwood	886.71	35	70.00
16	Anthony Horowitz	991.64	34	66.67
17	Michael Morpurgo	788.24	33	66.00
18	Suzanne Collins	848.82	32	64.00
19	James Patterson	982.79	31	63.27
20	Terry Pratchett	786.69	31	62.00
21	Stephenie Meyer	839.10	30	60.00
22	Lemony Snickett	843.84	30	60.00
23	Maya Angelou	918.55	28	56.00
24	Rick Riordan	806.96	26	52.00
25	J.D. Salinger	865.67	24	48.00

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
26	William Faulkner	942.66	24	48.00
27	Samuel Beckett	1005.36	23	46.00
28	Cassandra Clare	802.98	23	46.00
29	Veronica Roth	829.01	22	44.00
30	James Joyce	919.87	21	42.86
31	John Grisham	839.89	21	42.86
32	Haruki Murakami	877.13	21	42.00
33	Dan Brown	802.81	19	38.00
34	Ray Bradbury	1005.05	19	38.00
35	E.L. James	860.46	18	36.73
36	Rainbow Rowell	851.05	18	36.00
37	Jeff Kinney	925.37	18	36.00
38	Tom Clancy	857.59	18	36.00
39	Sarah J. Maas	754.44	17	34.69
40	Gabriel Garcia Marquez	1002.52	16	32.00
41	Thomas Wolfe	892.59	15	30.61
42	Danielle Steel	880.24	14	28.57
43	Ayn Rand	764.34	14	28.00
44	Vladimir Nabokov	929.16	14	28.00
45	Toni Morrison	906.16	13	26.00
46	Isaac Asimov	895.55	12	24.00
47	Kurt Vonnegut	825.59	11	22.00
48	Holly Black	839.35	11	22.00
49	Khaled Hosseini	859.31	10	20.00
50	E.B. White	922.38	9	18.00
51	Jack London	860.18	7	14.00
52	Hilary Mantel	831.21	7	14.00
53	Angie Thomas	890.46	7	14.00
54	Ralph Ellison	860.06	6	12.24
55	Lucinda Riley	820.87	6	12.00
56	Kate Dicamillo	819.36	4	8.16
57	Paulo Coelho	795.58	4	8.16
58	Karin Slaughter	846.64	3	6.00
59	Isabel Allende	855.72	3	6.00
60	Lisa Wingate	766.86	2	4.08

Table B.2: Chapter 3, English ART author selections by L1 UK English participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	J.K. Rowling	710.45	60	100.00
2	Roald Dahl	718.80	59	98.33
3	Agatha Christie	755.17	58	96.67
4	Stephen King	721.71	58	96.67
5	J.R.R. Tolkien	723.52	55	91.67
6	George Orwell	761.41	53	88.33
7	David Walliams	969.00	51	85.00
8	Ernest Hemingway	837.38	51	85.00
9	Dan Brown	806.02	49	81.67
10	F. Scott Fitzgerald	838.33	49	81.67
11	Jacqueline Wilson	849.05	49	81.67
12	Terry Pratchett	754.40	49	81.67
13	Virginia Woolf	825.08	49	81.67
14	Tom Clancy	885.97	48	80.00
15	Harper Lee	829.90	45	75.00
16	James Patterson	929.21	45	75.00
17	Danielle Steel	811.96	44	73.33
18	Anthony Horowitz	881.14	43	71.67
19	John Grisham	876.74	43	71.67
20	T.S. Eliot	827.73	43	71.67
21	Stephenie Meyer	930.31	39	65.00
22	Samuel Beckett	1021.62	37	61.67
23	James Joyce	816.67	35	58.33
24	Michael Morpurgo	829.92	33	55.00
25	Margaret Atwood	890.82	32	53.33
26	Suzanne Collins	1053.35	32	53.33
27	J.D. Salinger	871.20	31	51.67
28	Lemony Snickett	998.60	30	50.00
29	E.L. James	971.37	29	48.33
30	William Faulkner	1049.21	29	48.33
31	Isaac Asimov	918.02	28	46.67
32	Hilary Mantel	902.95	24	40.00
33	Maya Angelou	905.65	24	40.00

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
34	Ray Bradbury	863.06	24	40.00
35	Thomas Wolfe	965.24	23	38.33
36	Vladimir Nabokov	1042.88	21	35.00
37	John Green	867.43	19	31.67
38	E.B. White	950.17	17	28.33
39	Jack London	911.37	17	28.33
40	Kurt Vonnegut	868.58	17	28.33
41	Ayn Rand	790.43	15	25.00
42	Haruki Murakami	960.97	15	25.00
43	Karin Slaughter	865.93	15	25.00
44	Gabriel Garcia Marquez	1004.04	14	23.33
45	Toni Morrison	870.78	14	23.33
46	Rick Riordan	877.22	13	21.67
47	Veronica Roth	939.41	13	21.67
48	Cassandra Clare	881.99	10	16.67
49	Jeff Kinney	914.02	9	15.00
50	Khaled Hosseini	890.00	9	15.00
51	Paulo Coelho	852.21	8	13.33
52	Ralph Ellison	897.39	8	13.33
53	Sarah J. Maas	847.75	8	13.33
54	Isabel Allende	887.65	7	11.67
55	Lucinda Riley	939.02	6	10.00
56	Rainbow Rowell	916.76	6	10.00
57	Angie Thomas	934.87	4	6.67
58	Holly Black	834.60	4	6.67
59	Lisa Wingate	857.18	4	6.67
60	Kate Dicamillo	860.53	2	3.33

Table B.3: Chapter 4, English ART author selections by L1 English University of Oxford student participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	George Orwell	738.34	36	100.00
2	J.K. Rowling	646.73	36	100.00
3	Stephen King	746.67	36	100.00
4	Roald Dahl	730.49	35	97.22
5	Agatha Christie	781.24	34	94.44
6	J.R.R. Tolkien	748.17	34	94.44
7	F. Scott Fitzgerald	848.98	32	88.89
8	Jacqueline Wilson	791.85	32	88.89
9	T.S. Eliot	843.33	32	88.89
10	Virginia Woolf	749.95	32	88.89
11	David Walliams	832.73	30	83.33
12	Ernest Hemingway	883.41	30	83.33
13	John Green	787.02	30	83.33
14	Michael Morpurgo	870.44	30	83.33
15	Anthony Horowitz	918.99	29	80.56
16	Harper Lee	918.67	28	77.78
17	Stephenie Meyer	913.13	28	77.78
18	Lemony Snickett	989.38	27	75.00
19	Margaret Atwood	814.19	27	75.00
20	Maya Angelou	989.47	27	75.00
21	Suzanne Collins	849.89	26	72.22
22	Terry Pratchett	910.77	26	72.22
23	J.D. Salinger	992.68	24	66.67
24	Rick Riordan	878.81	23	63.89
25	William Faulkner	1081.12	22	61.11
26	Haruki Murakami	968.52	20	55.56
27	Veronica Roth	869.96	20	55.56
28	Dan Brown	1020.68	19	52.78
29	James Joyce	1087.35	19	52.78
30	James Patterson	1045.33	19	52.78
31	John Grisham	993.90	19	52.78
32	Samuel Beckett	1008.96	19	52.78
33	Cassandra Clare	849.12	16	44.44

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
34	Jeff Kinney	1011.81	16	44.44
35	Sarah J. Maas	901.29	15	41.67
36	Rainbow Rowell	893.72	13	36.11
37	Isaac Asimov	998.20	12	33.33
38	Khaled Hosseini	982.34	12	33.33
39	Tom Clancy	1058.08	12	33.33
40	Toni Morrison	976.56	12	33.33
41	E.L. James	870.44	11	30.56
42	Gabriel Garcia Marquez	1193.30	11	30.56
43	Holly Black	899.56	11	30.56
44	Ray Bradbury	951.77	11	30.56
45	Thomas Wolfe	1052.01	11	30.56
46	Vladimir Nabokov	1025.25	11	30.56
47	Ayn Rand	946.57	10	27.78
48	E.B. White	939.12	10	27.78
49	Hilary Mantel	859.33	10	27.78
50	Kurt Vonnegut	975.98	9	25.00
51	Danielle Steel	902.69	8	22.22
52	Jack London	912.39	7	19.44
53	Angie Thomas	977.21	6	16.67
54	Paulo Coelho	992.68	6	16.67
55	Isabel Allende	884.41	4	11.11
56	Karin Slaughter	969.44	4	11.11
57	Lucinda Riley	955.97	4	11.11
58	Kate Dicamillo	920.78	2	5.56
59	Ralph Ellison	1043.21	2	5.56
60	Lisa Wingate	866.18	1	2.78

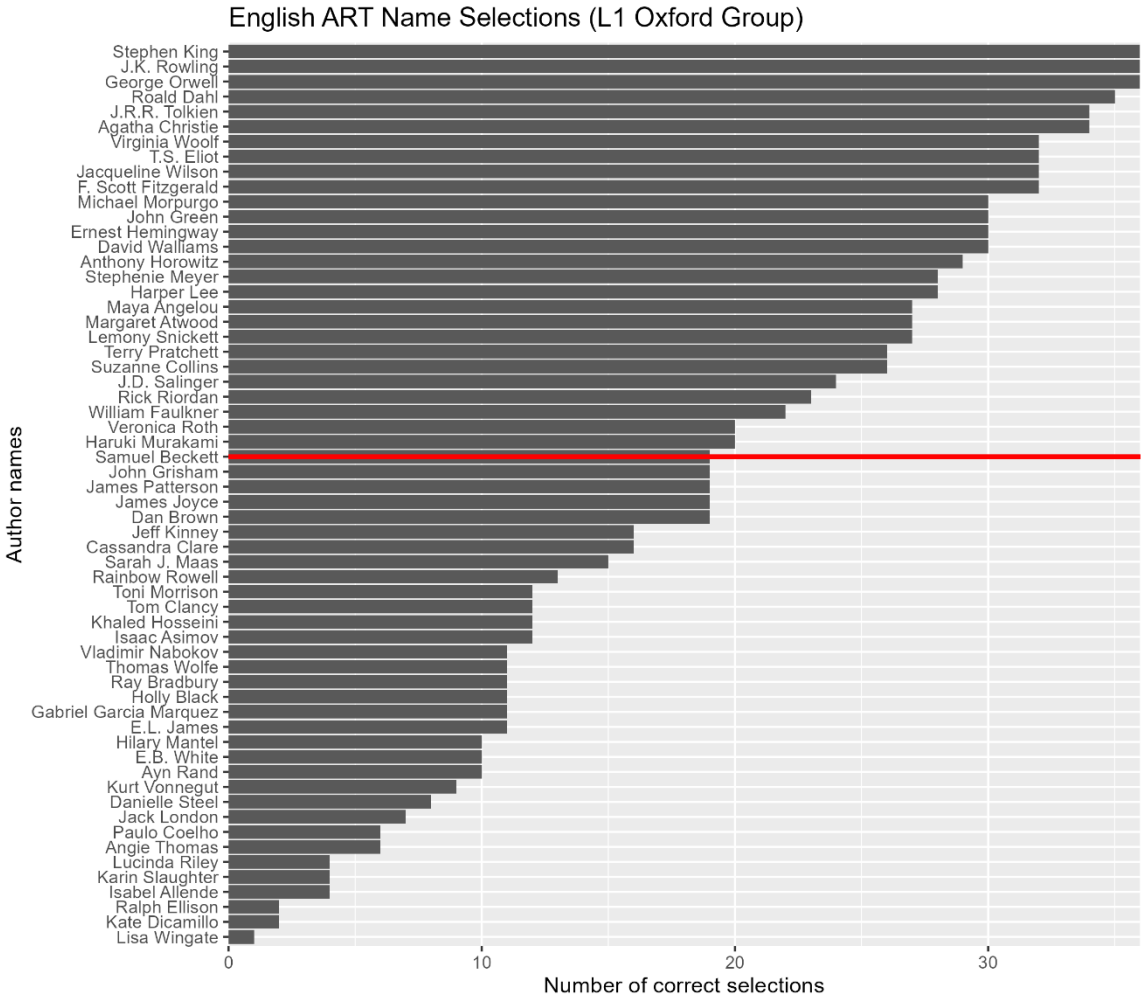


Figure B.1: Number of author name selections on the English ART, L1 English Oxford student participants (Chapter 4; total $n = 36$). Names at or above the red line (Samuel Beckett) are those with a recognition rate of 50% or greater.

Appendix B.1.2 English ARTs, L1 French Samples

Table B.4: Chapter 3, English ART author selections by L1 French participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	Stephen King	682.45	59	98.33
2	Agatha Christie	744.86	58	96.67
3	J.K. Rowling	695.24	58	96.67
4	George Orwell	838.61	57	95.00
5	J.R.R. Tolkien	737.47	53	88.33
6	Ernest Hemingway	847.44	46	76.67
7	F. Scott Fitzgerald	963.67	42	70.00
8	Samuel Beckett	931.04	42	70.00
9	Roald Dahl	816.26	40	66.67
10	Margaret Atwood	944.36	39	65.00
11	Virginia Woolf	784.93	39	65.00
12	Stephenie Meyer	947.31	36	60.00
13	Tom Clancy	885.29	36	60.00
14	Isaac Asimov	860.51	35	58.33
15	Jack London	843.83	35	58.33
16	Danielle Steel	879.67	33	55.00
17	Dan Brown	861.40	32	53.33
18	J.D. Salinger	851.01	31	51.67
19	James Joyce	900.41	31	51.67
20	Suzanne Collins	942.02	31	51.67
21	John Green	850.55	30	50.00
22	Paulo Coelho	813.48	29	48.33
23	Terry Pratchett	857.94	29	48.33
24	Harper Lee	900.98	28	46.67
25	Haruki Murakami	944.95	28	46.67
26	Ray Bradbury	917.81	28	46.67
27	E.L. James	864.12	27	45.00
28	Vladimir Nabokov	952.92	27	45.00
29	T.S. Eliot	856.49	25	41.67
30	William Faulkner	864.22	25	41.67
31	Gabriel Garcia Marquez	985.85	24	40.00

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
32	Toni Morrison	935.06	23	38.33
33	James Patterson	1030.20	22	36.67
34	Maya Angelou	920.61	21	35.00
35	John Grisham	953.34	20	33.33
36	Anthony Horowitz	971.72	17	28.33
37	Veronica Roth	917.90	16	26.67
38	Thomas Wolfe	1012.74	13	21.67
39	Isabel Allende	842.97	12	20.00
40	Sarah J. Maas	875.68	12	20.00
41	Lemony Snickett	842.53	11	18.33
42	Cassandra Clare	936.73	10	16.67
43	Holly Black	849.79	10	16.67
44	Kurt Vonnegut	785.96	9	15.00
45	Rainbow Rowell	945.28	9	15.00
46	Jeff Kinney	882.52	8	13.33
47	Rick Riordan	855.10	8	13.33
48	Karin Slaughter	909.35	7	11.67
49	Angie Thomas	899.11	6	10.00
50	Ayn Rand	801.09	6	10.00
51	E.B. White	804.73	6	10.00
52	Michael Morpurgo	889.79	6	10.00
53	Ralph Ellison	899.68	6	10.00
54	Hilary Mantel	882.83	4	6.67
55	Jacqueline Wilson	930.58	4	6.67
56	Lisa Wingate	851.88	4	6.67
57	David Walliams	872.60	2	3.33
58	Kate Dicamillo	898.38	2	3.33
59	Khaled Hosseini	855.36	1	1.67
60	Lucinda Riley	847.82	0	0.00

Table B.5: Chapter 5, English ART author selections by L1 French participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	Stephen King	704.75	51	100
2	Agatha Christie	751.45	50	98.04
3	J.K. Rowling	704.19	50	98.04
4	George Orwell	734.43	48	94.12
5	J.R.R. Tolkien	804.12	45	88.24
6	Ernest Hemingway	916.85	42	82.35
7	F. Scott Fitzgerald	1013.45	40	78.43
8	Samuel Beckett	879.12	40	78.43
9	Roald Dahl	800.63	34	66.67
10	Jack London	941.75	33	64.71
11	Virginia Woolf	838.84	33	64.71
12	Isaac Asimov	885.43	32	62.75
13	Dan Brown	864.87	31	60.78
14	Paulo Coelho	830.17	31	60.78
15	Tom Clancy	893.16	30	58.82
16	Margaret Atwood	1010.58	29	56.86
17	Suzanne Collins	955.18	29	56.86
18	James Patterson	916.07	28	54.90
19	John Green	870.02	28	54.90
20	James Joyce	925.78	27	52.94
21	Stephenie Meyer	1058.63	27	52.94
22	Danielle Steel	899.19	26	50.98
23	Ray Bradbury	823.19	26	50.98
24	E.L. James	902.55	25	49.02
25	Terry Pratchett	948.95	25	49.02
26	J.D. Salinger	869.24	24	47.06
27	Gabriel Garcia Marquez	1098.07	23	45.10
28	William Faulkner	913.42	23	45.10
29	Haruki Murakami	1109.47	21	41.18
30	Toni Morrison	946.99	21	41.18
31	Vladimir Nabokov	1112.44	21	41.18
32	Harper Lee	933.60	20	39.22

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
33	T.S. Eliot	904.63	20	39.22
34	John Grisham	864.87	16	31.37
35	Anthony Horowitz	1044.46	15	29.41
36	Isabel Allende	907.89	13	25.49
37	Veronica Roth	937.57	13	25.49
38	Lemony Snickett	854.73	11	21.57
39	Sarah J. Maas	930.55	11	21.57
40	Maya Angelou	902.31	10	19.61
41	Cassandra Clare	936.20	9	17.65
42	Thomas Wolfe	1005.67	9	17.65
43	E.B. White	881.29	7	13.73
44	Holly Black	822.19	7	13.73
45	Jeff Kinney	830.18	7	13.73
46	Angie Thomas	901.29	6	11.76
47	Ayn Rand	791.65	6	11.76
48	Jacqueline Wilson	1006.45	6	11.76
49	Lisa Wingate	849.72	6	11.76
50	Karin Slaughter	936.80	5	9.80
51	Kurt Vonnegut	850.61	5	9.80
52	Rainbow Rowell	883.5	5	9.80
53	David Walliams	884.07	4	7.84
54	Michael Morpurgo	902.90	4	7.84
55	Ralph Ellison	914.79	4	7.84
56	Rick Riordan	819.39	4	7.84
57	Khaled Hosseini	1021.39	3	5.88
58	Lucinda Riley	926.82	3	5.88
59	Hilary Mantel	958.27	2	3.92
60	Kate Dicamillo	905.51	2	3.92

Appendix B.1.3 English ARTs, L1 Korean Sample

Table B.6: Chapter 4, English ART author selections, L1 Korean participants.

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
1	J.K. Rowling	797.36	58	93.55
2	Ernest Hemingway	955.97	55	88.71
3	Stephen King	842.90	54	87.10
4	George Orwell	906.73	51	82.26
5	J.R.R. Tolkien	953.04	50	80.65
6	F. Scott Fitzgerald	960.09	42	67.74
7	Agatha Christie	936.41	39	62.90
8	Roald Dahl	896.67	39	62.90
9	Harper Lee	986.39	38	61.29
10	Haruki Murakami	1119.33	38	61.29
11	Virginia Woolf	890.56	36	58.06
12	T.S. Eliot	872.09	35	56.45
13	Dan Brown	834.83	32	51.61
14	James Patterson	1039.85	31	50.00
15	Margaret Atwood	999.40	30	48.39
16	J.D. Salinger	947.78	29	46.77
17	John Green	972.12	27	43.55
18	Stephenie Meyer	1075.84	27	43.55
19	John Grisham	909.85	26	41.94
20	Anthony Horowitz	1222.50	25	40.32
21	Tom Clancy	940.98	24	38.71
22	Paulo Coelho	934.42	23	37.10
23	Suzanne Collins	1001.07	23	37.10
24	E.B. White	923.71	22	35.48
25	Isaac Asimov	993.42	21	33.87
26	E.L. James	888.37	19	30.65
27	Ray Bradbury	840.82	19	30.65
28	James Joyce	942.39	18	29.03
29	Lemony Snickett	878.72	18	29.03
30	Maya Angelou	905.79	18	29.03
31	Rick Riordan	873.47	18	29.03
32	Samuel Beckett	967.09	17	27.42

Rank	Author Name	Mean Response Time (ms)	Selections	Percentage Correct
33	Thomas Wolfe	1059.06	16	25.81
34	William Faulkner	924.65	15	24.19
35	Jeff Kinney	898.18	14	22.58
36	Cassandra Clare	941.22	13	20.97
37	Gabriel Garcia Marquez	1204.66	12	19.35
38	Rainbow Rowell	899.90	12	19.35
39	Terry Pratchett	971.26	12	19.35
40	Veronica Roth	1018.88	12	19.35
41	Ayn Rand	885.06	11	17.74
42	Danielle Steel	942.88	11	17.74
43	David Walliams	1000.68	11	17.74
44	Kurt Vonnegut	862.31	11	17.74
45	Toni Morrison	952.10	11	17.74
46	Ralph Ellison	908.73	10	16.13
47	Vladimir Nabokov	1088.27	10	16.13
48	Hilary Mantel	959.06	8	12.90
49	Isabel Allende	1022.17	8	12.90
50	Jack London	946.64	8	12.90
51	Khaled Hosseini	1013.36	8	12.90
52	Sarah J. Maas	904.36	7	11.29
53	Jacqueline Wilson	1056.22	6	9.68
54	Karin Slaughter	952.61	6	9.68
55	Kate Dicamillo	925.25	6	9.68
56	Michael Morpurgo	914.59	6	9.68
57	Holly Black	804.84	5	8.06
58	Lisa Wingate	952.47	5	8.06
59	Angie Thomas	942.62	3	4.84
60	Lucinda Riley	836.94	3	4.84

Appendix B.1.4 French ARTs

Table B.7: Chapter 3, French ART author selections, L1 French participants.

Rank	Author Name	Nationality	Mean Response Time (ms)	Selections	Percentage Correct
1	Jean-Paul Sartre	French	797.50	57	95.00
2	Marc Levy	French	789.63	56	93.33
3	Guillaume Musso	French	866.35	42	70.00
4	Amélie Nothomb	French	806.90	41	68.33
5	Bernard Weber	French	862.94	36	60.00
6	Gérard De Villiers	French	1047.59	31	51.67
7	**Patricia Cornwell**	**Foreign**	917.60	31	51.67
8	**Tom Clancy**	**Foreign**	824.30	30	50.00
9	**Danielle Steel**	**Foreign**	897.93	28	46.67
10	**E. L. James**	**Foreign**	872.61	28	46.67
11	**Umberto Eco**	**Foreign**	837.24	27	45.00
12	Fred Vargas	French	855.33	26	43.33
13	**Ray Bradbury**	**Foreign**	906.06	25	41.67
14	**Gabriel García Márquez**	**Foreign**	1075.63	21	35.00
15	**Vladimir Nabokov**	**Foreign**	982.08	21	35.00
16	**Ken Follett**	**Foreign**	812.82	20	33.33
17	Maxime Chattam	French	818.00	20	33.33

Rank	Author Name	Nationality	Mean Response Time (ms)	Selections	Percentage Correct
18	Jean-Christophe Grangé	French	996.54	18	30.00
19	Patrick Modiano	French	979.79	17	28.33
20	**Toni Morrison**	**Foreign**	833.94	16	26.67
21	Jean-Christophe Rufin	French	1117.22	15	25.00
22	**Joyce Carol Oates**	**Foreign**	991.56	15	25.00
23	**T. C. Boyle**	**Foreign**	891.31	15	25.00
24	Virginie Grimaldi	French	1039.11	14	23.33
25	Delphine de Vigan	French	866.84	13	21.67
26	**Alexandre Soljenitsyne**	**Foreign**	903.07	11	18.33
27	**Isabel Allende**	**Foreign**	915.67	11	18.33
28	Joël Dicker	French	867.36	11	18.33
29	Édouard Louis	French	860.35	11	18.33
30	**Arnaldur Indridason**	**Foreign**	900.70	10	16.67
31	Claude Simon	French	866.85	10	16.67
32	**Raymond Chandler**	**Foreign**	969.33	10	16.67
33	**William Golding**	**Foreign**	989.08	10	16.67
34	**Günter Grass**	**Foreign**	807.77	8	13.33
35	Gilles Legardinier	French	938.64	7	11.67
36	Saint-John Perse	French	868.56	7	11.67

Rank	Author Name	Nationality	Mean Response Time (ms)	Selections	Percentage Correct
37	Valérie Perrin	French	929.89	7	11.67
38	**Martin Suter**	**Foreign**	811.34	5	8.33
39	**Miguel Ángel Asturias**	**Foreign**	937.95	5	8.33
40	Agnès Ledig	French	835.17	4	6.67

Appendix B.1.5 Korean ART

Table B.8: Chapter 4, Korean ART author selections, L1 Korean participants.

Rank	Author Name	English Translation	Nationality	Mean RT (ms)	Selections	Percentage Correct
1	어니스트 헤밍웨이	**Ernest Hemingway**	**Foreign**	999.33	57	91.94
2	제인 오스틴	**Jane Austen**	**Foreign**	998.09	55	88.71
3	J.R.R. 톨킨	**J.R.R. Tolkien**	**Foreign**	907.57	53	85.48
4	윤동주	Yoon Dong-ju	Korean	809.33	50	80.65
5	무라카미 하루키	**Haruki Murakami**	**Foreign**	975.03	45	72.58
6	베르나르 베르베르	**Bernard Werber**	**Foreign**	1011.42	45	72.58
7	애덤 스미스	**Adam Smith**	**Foreign**	1096.76	45	72.58
8	리처드 도킨스	**Richard Dawkins**	**Foreign**	1132.55	42	67.74
9	헤르만 헤세	**Hermann Hesse**	**Foreign**	1065.12	42	67.74
10	혜민 스님	Haemin Sunim	Korean	1190.00	42	67.74
11	김소월	Kim Sowol	Korean	861.88	40	64.52
12	요한 볼프강 폰 괴테	**Johann Wolfgang von Goethe**	**Foreign**	1423.47	39	62.90
13	한강	Han Kang	Korean	862.10	36	58.06
14	프란츠 카프카	**Franz Kafka**	**Foreign**	1057.76	33	53.23
15	칼 세이건	**Carl Sagan**	**Foreign**	1080.68	30	48.39
16	법륜	Pomnyun	Korean	1014.71	29	46.77

Rank	Author Name	English Translation	Nationality	Mean RT (ms)	Selections	Percentage Correct
17	알베르 카뮈	**Albert Camus**	**Foreign**	1138.25	29	46.77
18	히가시노 게이고	**Keigo Higashino**	**Foreign**	1084.83	29	46.77
19	마이클 샌델	**Michael Sandel**	**Foreign**	1077.98	28	45.16
20	백석	Baek Seok	Korean	936.29	28	45.16
21	신경숙	Shin Kyung-sook	Korean	849.70	28	45.16
22	알랭 드 보통	**Alain de Botton**	**Foreign**	1177.28	27	43.55
23	나태주	Na Tae-joo	Korean	1017.75	26	41.94
24	조정래	Jo Jung-rae	Korean	931.77	26	41.94
25	김영하	Kim Young-ha	Korean	895.29	24	38.71
26	신영복	Shin Young-bok	Korean	937.57	24	38.71
27	기욤 뮈소	**Guillaume Musso**	**Foreign**	1172.23	23	37.10
28	정유정	You-jeong Jeong	Korean	980.37	22	35.48
29	김미경	Kim Mi-kyung	Korean	1023.59	20	32.26
30	김난도	Kim Nan-do	Korean	905.56	19	30.65
31	미셸 푸코	**Michel Foucault**	**Foreign**	1088.52	19	30.65
32	오쿠다 히데오	**Hideo Okuda**	**Foreign**	1082.40	15	24.19
33	김진명	Kim Jin-myung	Korean	884.75	14	22.58
34	호메로스	**Homer**	**Foreign**	1142.50	14	22.58
35	더글라스 케네디	**Douglas Kennedy**	**Foreign**	1339.30	13	20.97

Rank	Author Name	English Translation	Nationality	Mean RT (ms)	Selections	Percentage Correct
36	요나스 요나손	**Jonas Jonasson**	**Foreign**	1152.33	11	17.74
37	박훈	Park Hoon	Korean	932.06	9	14.52
38	에쿠니 가오리	**Kaori Ekuni**	**Foreign**	1047.31	7	11.29
39	채사장	Chae Sa-jang	Korean	919.20	4	6.45
40	장 지글러	**Jean Ziegler**	**Foreign**	983.24	3	4.84

Appendix B.2 Author Fluency Tasks

Appendix B.2.1 English AFTs, L1 English Samples

Table B.9: Chapter 2, top 60 English Author Fluency Task name entries by L1 English Oxford University participants.

Rank	Author Name	Selections	Percentage
1	J.K. Rowling	40	80.00
2	William Shakespeare	28	56.00
3	J.R.R. Tolkien	19	38.00
4	Charles Dickens	18	36.00
5	Roald Dahl	17	34.00
6	Jane Austen	16	32.00
7	Jacqueline Wilson	15	30.00
8	Enid Blyton	11	22.00
9	Mary Shelley	10	20.00
10	Charlotte Brontë	9	18.00
11	C.S. Lewis	9	18.00
12	George Orwell	9	18.00
13	Edgar Allan Poe	8	16.00
14	Emily Brontë	8	16.00
15	John Green	8	16.00
16	Cassandra Clare	7	14.00
17	Michael Morpurgo	7	14.00
18	Oscar Wilde	7	14.00
19	Rick Riordan	7	14.00
20	Sally Rooney	7	14.00
21	Sarah J. Maas	7	14.00
22	Suzanne Collins	7	14.00
23	Agatha Christie	5	10.00
24	Ernest Hemingway	5	10.00
25	J.B. Priestley	5	10.00
26	Leigh Bardugo	5	10.00
27	Lewis Carroll	5	10.00

Rank	Author Name	Selections	Percentage
28	Neil Gaiman	5	10.00
29	Stephen King	5	10.00
30	Steven Pinker	5	10.00
31	Virginia Woolf	5	10.00
32	Harper Lee	4	8.00
33	Kazuo Ishiguro	4	8.00
34	Oliver Sacks	4	8.00
35	Philip Pullman	4	8.00
36	Terry Pratchett	4	8.00
37	Albert Camus	3	6.00
38	Arthur Conan Doyle	3	6.00
39	Arthur Miller	3	6.00
40	David Walliams	3	6.00
41	F. Scott Fitzgerald	3	6.00
42	Fyodor Dostoyevsky	3	6.00
43	Haruki Murakami	3	6.00
44	H.G. Wells	3	6.00
45	Holly Black	3	6.00
46	Lemony Snicket	3	6.00
47	Malorie Blackman	3	6.00
48	Margaret Atwood	3	6.00
49	Mark Twain	3	6.00
50	Maya Angelou	3	6.00
51	Patrick Ness	3	6.00
52	Percy Bysshe Shelley	3	6.00
53	Rainbow Rowell	3	6.00
54	Salman Rushdie	3	6.00
55	Stephenie Meyer	3	6.00
56	Veronica Roth	3	6.00
57	Yuval Noah Harari	3	6.00
58	Angela Carter	2	4.00
59	Anthony Horowitz	2	4.00
60	Beatrix Potter	2	4.00

Table B.10: Chapter 3, top 60 English Author Fluency Task name entries by L1 English participants.

Rank	Author Name	Selections	Percentage
1	J.K. Rowling	41	68.33
2	Charles Dickens	37	61.67
3	William Shakespeare	27	45.00
4	Enid Blyton	24	40.00
5	J.R.R. Tolkien	22	36.67
6	Stephen King	22	36.67
7	Roald Dahl	20	33.33
8	Jane Austen	18	30.00
9	Charlotte Brontë	15	25.00
10	Agatha Christie	14	23.33
11	Emily Brontë	14	23.33
12	Jacqueline Wilson	11	18.33
13	George Orwell	10	16.67
14	C.S. Lewis	9	15.00
15	Dan Brown	8	13.33
16	David Walliams	8	13.33
17	James Patterson	7	11.67
18	Arthur Conan Doyle	6	10.00
19	Lewis Carroll	6	10.00
20	Beatrix Potter	5	8.33
21	Colleen Hoover	5	8.33
22	Terry Pratchett	5	8.33
23	Catherine Cookson	4	6.67
24	Dean Koontz	4	6.67
25	F. Scott Fitzgerald	4	6.67
26	Harper Lee	4	6.67
27	H.G. Wells	4	6.67
28	Iain Banks	4	6.67
29	Ian Fleming	4	6.67
30	James Joyce	4	6.67
31	Jilly Cooper	4	6.67

Rank	Author Name	Selections	Percentage
32	Mary Shelley	4	6.67
33	Thomas Hardy	4	6.67
34	A.A. Milne	3	5.00
35	Danielle Steel	3	5.00
36	Dick Francis	3	5.00
37	Edgar Allan Poe	3	5.00
38	Ernest Hemingway	3	5.00
39	Hilary Mantel	3	5.00
40	Ian Rankin	3	5.00
41	Jackie Collins	3	5.00
42	James Herbert	3	5.00
43	J.D. Salinger	3	5.00
44	Jeffrey Archer	3	5.00
45	John Grisham	3	5.00
46	John le Carré	3	5.00
47	Judy Blume	3	5.00
48	Julia Donaldson	3	5.00
49	Lee Child	3	5.00
50	Leo Tolstoy	3	5.00
51	Lisa Jewell	3	5.00
52	Michael Morpurgo	3	5.00
53	Oscar Wilde	3	5.00
54	Peter James	3	5.00
55	Philip Pullman	3	5.00
56	Robert Harris	3	5.00
57	Rudyard Kipling	3	5.00
58	Suzanne Collins	3	5.00
59	T.S. Eliot	3	5.00
60	Ursula K. Le Guin	3	5.00

Table B.11: Chapter 4, top 60 English Author Fluency Task name entries by L1 English Oxford participants.

Rank	Author Name	Selections	Percentage
1	William Shakespeare	24	66.67
2	Jane Austen	23	63.89
3	J.K. Rowling	22	61.11
4	Charles Dickens	20	55.56
5	J.R.R. Tolkien	19	52.78
6	Charlotte Brontë	14	38.89
7	Emily Brontë	14	38.89
8	George Orwell	14	38.89
9	Roald Dahl	11	30.56
10	Jacqueline Wilson	10	27.78
11	Colleen Hoover	8	22.22
12	Enid Blyton	7	19.44
13	Stephen King	7	19.44
14	Suzanne Collins	7	19.44
15	C.S. Lewis	6	16.67
16	Agatha Christie	5	13.89
17	Haruki Murakami	5	13.89
18	Margaret Atwood	5	13.89
19	Michael Morpurgo	5	13.89
20	Oscar Wilde	5	13.89
21	Sally Rooney	5	13.89
22	Stephenie Meyer	5	13.89
23	Terry Pratchett	5	13.89
24	F. Scott Fitzgerald	4	11.11
25	Mary Shelley	4	11.11
26	Sarah J. Maas	4	11.11
27	Anne Brontë	3	8.33
28	Cathy Cassidy	3	8.33
29	David Walliams	3	8.33
30	Edgar Allan Poe	3	8.33
31	Friedrich Nietzsche	3	8.33

Rank	Author Name	Selections	Percentage
32	George R.R. Martin	3	8.33
33	James Patterson	3	8.33
34	Julia Donaldson	3	8.33
35	Kazuo Ishiguro	3	8.33
36	Leigh Bardugo	3	8.33
37	Leo Tolstoy	3	8.33
38	Malorie Blackman	3	8.33
39	Patrick Ness	3	8.33
40	Rick Riordan	3	8.33
41	Sylvia Plath	3	8.33
42	Taylor Jenkins Reid	3	8.33
43	T.S. Eliot	3	8.33
44	Ursula K. Le Guin	3	8.33
45	Virginia Woolf	3	8.33
46	Alice Oseman	2	5.56
47	Anthony Horowitz	2	5.56
48	Arthur Conan Doyle	2	5.56
49	Arthur Miller	2	5.56
50	Bill Bryson	2	5.56
51	Douglas Adams	2	5.56
52	Dr. Seuss	2	5.56
53	Emily Henry	2	5.56
54	Erin Hunter	2	5.56
55	Ernest Hemingway	2	5.56
56	Harper Lee	2	5.56
57	Holly Black	2	5.56
58	Holly Smale	2	5.56
59	Ian Fleming	2	5.56
60	Ian McEwan	2	5.56

Table B.12: Chapter 5, top 60 English Author Fluency Task name entries by L1 English participants.

Rank	Author Name	Selections	Percentage
1	William Shakespeare	29	58.00
2	J.K. Rowling	28	56.00
3	Stephen King	22	44.00
4	Charles Dickens	18	36.00
5	J.R.R. Tolkien	17	34.00
6	Roald Dahl	15	30.00
7	Enid Blyton	15	30.00
8	Jane Austen	10	20.00
9	Agatha Christie	10	20.00
10	George Orwell	9	18.00
11	Jacqueline Wilson	8	16.00
12	David Walliams	8	16.00
13	C.S. Lewis	8	16.00
14	Ernest Hemingway	7	14.00
15	Salman Rushdie	6	12.00
16	Ian Fleming	6	12.00
17	Dan Brown	5	10.00
18	Emily Brontë	5	10.00
19	Mark Twain	5	10.00
20	Terry Pratchett	4	8.00
21	Margaret Atwood	4	8.00
22	John Grisham	4	8.00
23	Leo Tolstoy	4	8.00
24	George Eliot	4	8.00
25	Charlotte Brontë	3	6.00
26	Lewis Carroll	3	6.00
27	Dean Koontz	3	6.00
28	Jeffrey Archer	3	6.00
29	Hunter S. Thompson	3	6.00
30	John Steinbeck	3	6.00
31	Michael Crichton	3	6.00

Rank	Author Name	Selections	Percentage
32	Sally Rooney	3	6.00
33	Jules Verne	3	6.00
34	Julia Donaldson	3	6.00
35	Lee Child	3	6.00
36	Zadie Smith	3	6.00
37	J.D. Salinger	2	4.00
38	Oscar Wilde	2	4.00
39	George R.R. Martin	2	4.00
40	Arthur Conan Doyle	2	4.00
41	Hilary Mantel	2	4.00
42	Jackie Collins	2	4.00
43	James Herbert	2	4.00
44	James Patterson	2	4.00
45	Lisa Jewell	2	4.00
46	Carl Hiaasen	2	4.00
47	Dr. Seuss	2	4.00
48	Franz Kafka	2	4.00
49	Geoffrey Chaucer	2	4.00
50	Jojo Moyes	2	4.00
51	Maeve Binchy	2	4.00
52	Marian Keyes	2	4.00
53	Philip K. Dick	2	4.00
54	Poppy Z. Brite	2	4.00
55	Sophie Kinsella	2	4.00
56	Virginia Woolf	2	4.00
57	Carl Sagan	2	4.00
58	Catherine Cookson	2	4.00
59	Edgar Allan Poe	2	4.00
60	Evelyn Waugh	2	4.00

Appendix B.2.2 English AFT, L1 French Samples

Table B.13: Chapter 3, top 60 English Author Fluency Task name selections, L1 French group.

Rank	Author Name	Nationality	Selections	Percentage
1	J.K. Rowling	Foreign	38	63.33
2	William Shakespeare	Foreign	35	58.33
3	Agatha Christie	Foreign	24	40.00
4	J.R.R. Tolkien	Foreign	24	40.00
5	Stephen King	Foreign	19	31.67
6	Jane Austen	Foreign	16	26.67
7	Edgar Allan Poe	Foreign	14	23.33
8	George Orwell	Foreign	13	21.67
9	Charles Dickens	Foreign	12	20.00
10	Ernest Hemingway	Foreign	12	20.00
11	Emily Brontë	Foreign	11	18.33
12	Roald Dahl	Foreign	11	18.33
13	C.S. Lewis	Foreign	9	15.00
14	Arthur Conan Doyle	Foreign	8	13.33
15	Charlotte Brontë	Foreign	7	11.67
16	George R.R. Martin	Foreign	6	10.00
17	Neil Gaiman	Foreign	6	10.00
18	Oscar Wilde	Foreign	6	10.00
19	Lewis Carroll	Foreign	5	8.33
20	Philip K. Dick	Foreign	5	8.33
21	Stephenie Meyer	Foreign	5	8.33
22	Virginia Woolf	Foreign	5	8.33
23	Aldous Huxley	Foreign	4	6.67
24	F. Scott Fitzgerald	Foreign	4	6.67
25	Isaac Asimov	Foreign	4	6.67
26	John Green	Foreign	4	6.67
27	Leigh Bardugo	Foreign	4	6.67
28	Terry Pratchett	Foreign	4	6.67
29	Anne Brontë	Foreign	3	5.00

Rank	Author Name	Nationality	Selections	Percentage
30	Diana Wynne Jones	Foreign	3	5.00
31	Emily Dickinson	Foreign	3	5.00
32	H.P. Lovecraft	Foreign	3	5.00
33	James Joyce	Foreign	3	5.00
34	Mary Shelley	Foreign	3	5.00
35	Salman Rushdie	Foreign	3	5.00
36	Suzanne Collins	Foreign	3	5.00
37	Anne Rice	Foreign	2	3.33
38	Bram Stoker	Foreign	2	3.33
39	Brontë Sisters	Foreign	2	3.33
40	Cassandra Clare	Foreign	2	3.33
41	**Charles Baudelaire**	**French**	2	3.33
42	Dale Carnegie	Foreign	2	3.33
43	Dan Brown	Foreign	2	3.33
44	Dan Simmons	Foreign	2	3.33
45	Danielle Steel	Foreign	2	3.33
46	Donna Tartt	Foreign	2	3.33
47	E.L. James	Foreign	2	3.33
48	Frank Herbert	Foreign	2	3.33
49	Friedrich Nietzsche	Foreign	2	3.33
50	Gillian Flynn	Foreign	2	3.33
51	H.G. Wells	Foreign	2	3.33
52	Holly Black	Foreign	2	3.33
53	Jack London	Foreign	2	3.33
54	James Baldwin	Foreign	2	3.33
55	John Milton	Foreign	2	3.33
56	John Steinbeck	Foreign	2	3.33
57	Margaret Atwood	Foreign	2	3.33
58	Nathaniel Hawthorne	Foreign	2	3.33
59	Patricia Cornwell	Foreign	2	3.33
60	Paul Auster	Foreign	2	3.33

Table B.14: Chapter 5, top 60 English Author Fluency Task name selections, L1 French group.

Rank	Author Name	Nationality	Selections	Percentage
1	William Shakespeare	Foreign	32	64.00
2	J.K. Rowling	Foreign	28	56.00
3	Agatha Christie	Foreign	20	40.00
4	Stephen King	Foreign	20	40.00
5	J.R.R. Tolkien	Foreign	18	36.00
6	Jane Austen	Foreign	9	18.00
7	Charles Dickens	Foreign	8	16.00
8	George Orwell	Foreign	7	14.00
9	Isaac Asimov	Foreign	7	14.00
10	Arthur Conan Doyle	Foreign	6	12.00
11	Roald Dahl	Foreign	6	12.00
12	Philip K. Dick	Foreign	6	12.00
13	**Victor Hugo**	**French**	6	12.00
14	Edgar Allan Poe	Foreign	5	10.00
15	George R.R. Martin	Foreign	5	10.00
16	Paul Auster	Foreign	5	10.00
17	Ray Bradbury	Foreign	5	10.00
18	Oscar Wilde	Foreign	5	10.00
19	Emily Brontë	Foreign	4	8.00
20	Mary Higgins Clark	Foreign	4	8.00
21	Ernest Hemingway	Foreign	3	6.00
22	Lewis Carroll	Foreign	3	6.00
23	Sylvia Plath	Foreign	3	6.00
24	Terry Pratchett	Foreign	3	6.00
25	Emily Dickinson	Foreign	3	6.00
26	H.P. Lovecraft	Foreign	3	6.00
27	Michael Crichton	Foreign	3	6.00
28	Neil Gaiman	Foreign	3	6.00
29	**Voltaire**	**French**	3	6.00
30	**Albert Camus**	**French**	3	6.00
31	Danielle Steel	Foreign	3	6.00

Rank	Author Name	Nationality	Selections	Percentage
32	Donna Tartt	Foreign	3	6.00
33	**Gustave Flaubert**	**French**	3	6.00
34	James Ellroy	Foreign	3	6.00
35	Stephen Hawking	Foreign	3	6.00
36	**Émile Zola**	**French**	3	6.00
37	**Jean-Paul Sartre**	**French**	3	6.00
38	Dan Brown	Foreign	2	4.00
39	Charlotte Brontë	Foreign	2	4.00
40	Suzanne Collins	Foreign	2	4.00
41	C.S. Lewis	Foreign	2	4.00
42	Harlan Coben	Foreign	2	4.00
43	Jack London	Foreign	2	4.00
44	Dale Carnegie	Foreign	2	4.00
45	**Fred Vargas**	**French**	2	4.00
46	Fyodor Dostoyevsky	Foreign	2	4.00
47	Henry David Thoreau	Foreign	2	4.00
48	Karl Marx	Foreign	2	4.00
49	Margaret Atwood	Foreign	2	4.00
50	**Stendhal**	**French**	2	4.00
51	Tom Clancy	Foreign	2	4.00
52	Baruch Spinoza	Foreign	2	4.00
53	Brontë Sisters	Foreign	2	4.00
54	Don DeLillo	Foreign	2	4.00
55	Frank Herbert	Foreign	2	4.00
56	Franz Kafka	Foreign	2	4.00
57	Gillian Flynn	Foreign	2	4.00
58	**Guillaume Musso**	**French**	2	4.00
59	Immanuel Kant	Foreign	2	4.00
60	Lucy Foley	Foreign	2	4.00

Appendix B.2.3 English AFT, L1 Korean Sample

Table B.15: Chapter 4, top 60 English Author Fluency Task name selections by L1 Korean participants.

Rank	Author Name	Selections	Percentage
1	J.K. Rowling	48	80.00
2	William Shakespeare	27	45.00
3	J.R.R. Tolkien	18	30.00
4	George Orwell	17	28.33
5	Agatha Christie	16	26.67
6	Ernest Hemingway	16	26.67
7	Jane Austen	16	26.67
8	Stephen King	16	26.67
9	Charles Dickens	15	25.00
10	Arthur Conan Doyle	10	16.67
11	Roald Dahl	10	16.67
12	Dan Brown	8	13.33
13	Dr. Seuss	7	11.67
14	Edgar Allan Poe	7	11.67
15	F. Scott Fitzgerald	7	11.67
16	George R.R. Martin	6	10.00
17	Leo Tolstoy	6	10.00
18	Mark Twain	6	10.00
19	Virginia Woolf	6	10.00
20	Charlotte Brontë	5	8.33
21	C.S. Lewis	5	8.33
22	Franz Kafka	5	8.33
23	Lewis Carroll	5	8.33
24	Rick Riordan	5	8.33
25	Carl Sagan	4	6.67
26	Colleen Hoover	4	6.67
27	Emily Brontë	4	6.67
28	Harper Lee	4	6.67
29	Hermann Hesse	4	6.67

Rank	Author Name	Selections	Percentage
30	Yuval Noah Harari	4	6.67
31	Emily Dickinson	3	5.00
32	Mary Shelley	3	5.00
33	Richard Dawkins	3	5.00
34	Shel Silverstein	3	5.00
35	Stephenie Meyer	3	5.00
36	Suzanne Collins	3	5.00
37	Albert Camus	2	3.33
38	Audre Lorde	2	3.33
39	Bernard Werber	2	3.33
40	Dale Carnegie	2	3.33
41	Han Kang	2	3.33
42	Henry David Thoreau	2	3.33
43	Ian McEwan	2	3.33
44	Malcolm Gladwell	2	3.33
45	Margaret Atwood	2	3.33
46	Maya Angelou	2	3.33
47	Min Jin Lee	2	3.33
48	Nathaniel Hawthorne	2	3.33
49	Oscar Wilde	2	3.33
50	Samuel Beckett	2	3.33
51	Sigmund Freud	2	3.33
52	Walt Whitman	2	3.33
53	William Golding	2	3.33
54	Adam Kay	1	1.67
55	Alan Turing	1	1.67
56	Alexander McCall Smith	1	1.67
57	Alice Walker	1	1.67
58	Alyssa Cole	1	1.67
59	Amor Towles	1	1.67
60	Angela Davis	1	1.67

Appendix B.2.4 French AFT

Table B.16: Chapter 3, top 60 French Author Fluency Task name selections, L1 French group. "Language" = primary language of publication.

Rank	Author Name	Language	Selections	Percentage
1	Victor Hugo	French	42	70.00
2	Émile Zola	French	33	55.00
3	Molière	French	30	50.00
4	Voltaire	French	24	40.00
5	Albert Camus	French	19	31.67
6	Charles Baudelaire	French	19	31.67
7	Guy de Maupassant	French	17	28.33
8	Honoré de Balzac	French	17	28.33
9	Jean-Jacques Rousseau	French	17	28.33
10	Jean Racine	French	15	25.00
11	Jean de La Fontaine	French	14	23.33
12	Gustave Flaubert	French	13	21.67
13	Marcel Proust	French	12	20.00
14	Arthur Rimbaud	French	11	18.33
15	Pierre Corneille	French	11	18.33
16	Amélie Nothomb	French	11	18.33
17	George Sand	French	11	18.33
18	Alexandre Dumas	French	10	16.67
19	Jules Verne	French	10	16.67
20	Marc Levy	French	10	16.67
21	Guillaume Musso	French	9	15.00
22	Jean-Paul Sartre	French	9	15.00
23	Louis-Ferdinand Céline	French	9	15.00
24	Paul Verlaine	French	9	15.00

Rank	Author Name	Language	Selections	Percentage
25	Simone De Beauvoir	French	9	15.00
26	Stendhal	French	9	15.00
27	Antoine De Saint-Exupéry	French	7	11.67
28	**JK Rowling**	**Foreign**	7	11.67
29	Michel Houellebecq	French	7	11.67
30	François Rabelais	French	6	10.00
31	La Comtesse de Ségur	French	6	10.00
32	Colette	French	6	10.00
33	Alfred de Musset	French	6	10.00
34	Guillaume Apollinaire	French	4	6.67
35	Virginie Despentes	French	4	6.67
36	**Roald Dahl**	**Foreign**	4	6.67
37	Marguerite Duras	French	4	6.67
38	Marcel Pagnol	French	4	6.67
39	Montesquieu	French	4	6.67
40	Edmond Rostand	French	4	6.67
41	Boris Vian	French	4	6.67
42	Alphonse de Lamartine	French	4	6.67
43	**Agatha Christie**	**Foreign**	4	6.67
44	Hervé Bazin	French	3	5.00
45	Paul Éluard	French	3	5.00
46	Marguerite Yourcenar	French	3	5.00
47	Madame de Lafayette	French	3	5.00
48	**JRR Tolkien**	**Foreign**	3	5.00
49	Jean Anouilh	French	3	5.00
50	Eugène Ionesco	French	3	5.00
51	**CS Lewis**	**Foreign**	3	5.00
52	Frédéric Beigbeder	French	3	5.00

Rank	Author Name	Language	Selections	Percentage
53	Fred Vargas	French	3	5.00
54	Denis Diderot	French	3	5.00
55	Charles Perrault	French	3	5.00
56	Bernard Werber	French	3	5.00
57	Ronsard	French	3	5.00
58	Pierre Choderlos de Laclos	French	3	5.00
59	Simone Veil	French	3	5.00
60	**Stephen King**	**Foreign**	3	5.00

Appendix B.2.5 Korean AFT

Table B.17: Chapter 4, top 60 Korean Author Fluency Task name selections, L1 Korean group. "Language" = primary language of publication.

Rank	Author Name	English Translation	Language	Selections	Percentage
1	한강	Han Kang	Korean	17	27.42
2	이상	Lee Sang	Korean	13	20.97
3	윤동주	Yoon Dong-ju	Korean	11	17.74
4	박완서	Park Wan-suh	Korean	10	16.13
5	J.K. 롤링	**J.K. Rowling**	**Foreign**	9	14.52
6	공지영	Gong Ji-young	Korean	9	14.52
7	김영하	Kim Young-ha	Korean	8	12.90
8	박경리	Park Kyung-ni	Korean	7	11.29
9	신경숙	Shin Kyung-sook	Korean	7	11.29
10	이문열	Lee Moon-yeol	Korean	7	11.29

Rank	Author Name	English Translation	Language	Selections	Percentage
11	J.R.R. 톨킨	**J.R.R. Tolkien**	**Foreign**	6	9.68
12	김소월	Kim Sowol	Korean	6	9.68
13	이외수	Lee Oi-soo	Korean	6	9.68
14	김유정	Kim Yoo-jung	Korean	5	8.06
15	베르나르 베르베르	**Bernard Werber**	**Foreign**	5	8.06
16	이민진	Min Jin Lee	Korean	5	8.06
17	최은영	Choi Eun-young	Korean	5	8.06
18	구병모	Gu Byeong-mo	Korean	4	6.45
19	김훈	Kim Hoon	Korean	4	6.45
20	나혜석	Na Hye-sok	Korean	4	6.45
21	류시화	Ryu Shi-hwa	Korean	4	6.45
22	무라카미 하루키	**Haruki Murakami**	**Foreign**	4	6.45
23	조남주	Cho Nam-joo	Korean	4	6.45
24	김초엽	Kim Cho-yeop	Korean	3	4.84
25	어니스트 헤밍웨이	**Ernest Hemingway**	**Foreign**	3	4.84
26	유시민	Yoo Si-min	Korean	3	4.84
27	정재승	Jeong Jae-seung	Korean	3	4.84
28	조지 오웰	**George Orwell**	**Foreign**	3	4.84
29	황석영	Hwang Sok-yong	Korean	3	4.84
30	김금희	Kim Geum-hee	Korean	2	3.23
31	김동리	Kim Dong-ni	Korean	2	3.23

Rank	Author Name	English Translation	Language	Selections	Percentage
32	김동인	Kim Dong-in	Korean	2	3.23
33	김만중	Kim Man-jung	Korean	2	3.23
34	김미경	Kim Mi-kyung	Korean	2	3.23
35	김은숙	Kim Eun-sook	Korean	2	3.23
36	김진명	Kim Jin-myung	Korean	2	3.23
37	나태주	Na Tae-joo	Korean	2	3.23
38	댄 브라운	**Dan Brown**	**Foreign**	2	3.23
39	레프 톨스토이	**Leo Tolstoy**	**Foreign**	2	3.23
40	마광수	Ma Kwang-soo	Korean	2	3.23
41	무라카미 류	**Ryu Murakami**	**Foreign**	2	3.23
42	박지원	Park Ji-won	Korean	2	3.23
43	백남준	Nam June Paik	Korean	2	3.23
44	버지니아 울프	**Virginia Woolf**	**Foreign**	2	3.23
45	서정주	Seo Jeong-ju	Korean	2	3.23
46	셰익스피어	**William Shakespeare**	**Foreign**	2	3.23
47	스티븐 킹	**Stephen King**	**Foreign**	2	3.23
48	알베르 까뮈	**Albert Camus**	**Foreign**	2	3.23
49	오정희	Oh Jung-hee	Korean	2	3.23
50	이광수	Lee Kwang-soo	Korean	2	3.23

Rank	Author Name	English Translation	Language	Selections	Percentage
51	이석원	Lee Seok-won	Korean	2	3.23
52	정약용	Jeong Yak-yong	Korean	2	3.23
53	정유정	Jeong You-jeong	Korean	2	3.23
54	정호승	Jeong Ho-seung	Korean	2	3.23
55	제인 오스틴	**Jane Austen**	**Foreign**	2	3.23
56	조정래	Cho Jung-rae	Korean	2	3.23
57	피천득	Pi Cheon-deuk	Korean	2	3.23
58	허준	Heo Jun	Korean	2	3.23
59	현진건	Hyun Jin-geon	Korean	2	3.23
60	황보름	Hwang Boreum	Korean	2	3.23

Appendix B.3 Lexical decision task for fiction keywords

Table B.18: Lexical decision task using keywords of literary fiction, L1 English Oxford University students (Chapter 2).

Note: Stimuli arranged by word/non-word and alphabetically, with average response times in milliseconds (ms), and percentage of correct selections. Matching non-words generated from Wuggy software (Keuleers & Brysbaert, 2010). POS = part of speech; prev. = prevalence norms; prev. quant = prevalence ratings as quantiles (1 = lowest, 5 = highest), Log Ratio = fiction keyness rating.

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
aspirant	word	8	adj.	0.938	3	1.02112	76	962.51
bawdy	word	5	adj.	1.210	5	1.037766	40	884.09
bereft	word	6	adj.	1.116	4	1.06461	60	831.30
bestial	word	7	adj.	1.174	5	1.081971	76	836.96
clamber	word	7	verb	1.036	4	1.073795	86	867.15
coffer	word	6	noun	1.058	4	0.991935	38	877.72
defrock	word	7	verb	1.221	5	1.019224	48	855.13

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
denizen	word	7	noun	0.689	1	1.03838	36	771.71
derision	word	8	noun	1.027	4	1.03436	56	867.50
disquiet	word	8	noun	1.097	4	1.02841	82	858.65
epicure	word	7	noun	1.060	4	1.032017	72	976.92
estimable	word	9	adj.	0.960	3	1.056108	82	896.72
exult	word	5	verb	0.897	3	1.034015	58	876.97
feckless	word	8	adj.	0.842	2	1.082828	66	866.85
functionary	word	11	noun	1.196	5	1.05709	90	888.42
genteel	word	7	adj.	0.839	2	0.99994	34	901.47
gristly	word	7	adj.	0.635	1	1.024446	72	1007.29
gumshoe	word	7	noun	0.795	2	1.073701	40	974.79
hackneyed	word	9	adj.	0.847	2	1.087996	38	1046.24
halitosis	word	9	noun	1.190	5	1.043688	40	928.24

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
hangar	word	6	noun	1.167	5	1.083413	44	922.27
hellcat	word	7	noun	1.053	4	1.036781	44	912.23
heraldic	word	8	adj.	0.829	2	1.002165	70	844.43
hothouse	word	8	noun	1.194	5	1.099792	58	977.92
ignominiously	word	13	adv.	0.844	2	1.085382	44	1162.43
imbecility	word	10	noun	0.945	3	1.045219	78	1030.63
incommunicado	word	13	adj.	0.939	3	1.087454	28	1252.95
lackey	word	6	noun	1.216	5	1.072472	54	907.99
lichen	word	6	noun	1.173	5	1.057287	68	880.49
longshoreman	word	12	noun	1.141	5	0.995806	48	1150.08
luridly	word	7	adv.	0.953	3	1.006644	58	992.44
madcap	word	6	adj.	0.801	2	1.013746	40	979.17
malady	word	6	noun	1.084	4	1.056113	54	929.24

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
misspent	word	8	verb	1.074	4	1.074966	78	890.67
nautilus	word	8	noun	1.060	4	1.019418	26	882.38
necromancer	word	11	noun	0.970	3	1.01367	76	891.29
ornery	word	6	adj.	0.927	3	1.084818	30	885.48
ostentation	word	11	noun	1.170	5	1.025032	82	964.94
pique	word	5	noun / verb	1.163	5	1.030216	74	857.96
plebeian	word	8	adj. / noun	1.030	4	1.05056	62	884.97
prepossess	word	10	verb	0.868	2	1.089418	72	1185.80
resplendent	word	11	adj.	1.007	3	1.094988	48	996.22
rummy	word	5	noun	1.159	5	1.060364	56	884.12
smoky	word	5	adj.	1.203	5	1.035985	94	731.83
snugly	word	6	adv.	1.205	5	0.991433	62	893.24
thrall	word	6	noun	0.996	3	1.033626	68	827.75

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
totter	word	6	verb	0.896	3	1.039218	62	946.21
unexceptionable	word	15	adj.	0.770	1	1.081049	90	1040.26
venturesome	word	11	adj.	1.196	5	1.066198	76	1030.17
atcarant	non-word	8					6	927.99
autibible	non-word	9					6	1163.88
balicofos	non-word	9					0	785.51
beripe	non-word	6					12	819.55
blistry	non-word	7					32	945.11
busteil	non-word	7					12	797.90
cauly	non-word	5					2	837.58
clebeood	non-word	8					0	763.96
coccer	non-word	6					4	778.24
dedrack	non-word	7					2	829.56

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
denwriet	non-word	8					0	719.19
ecicose	non-word	7					0	822.09
emuor	non-word	5					2	793.05
fecalpic	non-word	8					2	870.34
fibren	non-word	6					16	887.29
fuledly	non-word	7					8	995.35
futter	non-word	6					28	937.99
gendiel	non-word	7					8	782.78
gluky	non-word	5					4	709.04
gumstoo	non-word	7					0	788.59
hancor	non-word	6					14	860.89
hesslat	non-word	7					2	730.26
hingsionary	non-word	11					8	949.66

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
hophound	non-word	8					14	988.50
illefarity	non-word	10					4	890.15
inveggenicamo	non-word	13					0	981.38
iprilodiously	non-word	13					6	952.57
lamnomancer	non-word	11					6	1129.37
lereknoreman	non-word	12					0	890.98
lickem	non-word	6					2	850.11
madtop	non-word	6					30	970.53
misspine	non-word	8					26	1016.04
orogmoptionable	non-word	15					4	1149.19
poque	non-word	5					10	815.18
prepirress	non-word	10					4	923.47
quogly	non-word	6					0	725.05

Stimulus	Answer	Word Length	POS	prev.	prev. quant	Log Ratio	Selec. (%)	Mean RTs (ms)
rephrastent	non-word	11					12	954.34
rurry	non-word	5					8	873.45
spimber	non-word	7					2	832.48
sunady	non-word	6					0	967.52
sunipen	non-word	7					0	806.09
threll	non-word	6					10	781.25
tuckless	non-word	8					56	898.37
uncery	non-word	6					8	805.11
uttincation	non-word	11					6	972.83
ventumpmime	non-word	11					2	994.32
wiltilus	non-word	8					0	768.20

Appendix B.4 LexTALEs

Appendix B.4.1 English LexTALE Selections and Response Times, all cohorts (Chapters 3 & 4)

Table B.19: English LexTALE stimuli arranged by word/non-word and alphabetically, with average response times (RT) in milliseconds (ms), and percentage of selections for each cohort.

Note: EN = L1 UK English; EN (OX) = L1 English (Oxford) students, FR = L1 French / L2 English participants; KR = L1 Korean / L2 English participants.

			EN		EN (OX)		FR		KR	
	Stimulus	***Answer***	***Mean RT***	*%*	***Mean RT***	*%*	***Mean RT***	*%*	***Mean RT***	*%*
1	**ablaze**	word	849.83	95.00	841.03	97.22	901.21	48.33	1037.03	58.06
2	**allied**	word	768.22	96.67	742.24	100.00	892.50	85.00	933.55	88.71
3	**bewitch**	word	810.25	98.33	723.33	100.00	911.42	71.67	1028.77	85.48
4	**breeding**	word	731.93	98.33	675.15	100.00	819.20	95.00	832.43	100.00
5	**carbohydrate**	word	950.66	100.00	864.66	100.00	1084.24	96.67	1202.08	95.16
6	**celestial**	word	736.02	96.67	749.78	100.00	825.71	86.67	1065.20	88.71
7	**censorship**	word	793.70	98.33	733.53	100.00	829.38	90.00	934.72	93.55
8	**cleanliness**	word	731.85	98.33	727.68	97.22	928.05	91.67	1204.67	82.26
9	**cylinder**	word	694.51	96.67	720.25	100.00	816.82	93.33	843.39	98.39
10	**dispatch**	word	710.47	100.00	759.48	100.00	697.94	98.33	716.03	100.00
11	**eloquence**	word	926.90	88.33	930.83	100.00	793.37	98.33	1652.23	75.81
12	**festivity**	word	843.20	100.00	792.34	97.22	721.82	100.00	1031.03	91.94

			EN		EN (OX)		FR		KR	
	Stimulus	*Answer*	*Mean RT*	*%*	*Mean RT*	*%*	*Mean RT*	*%*	*Mean RT*	*%*
13	**flaw**	word	675.59	96.67	713.62	100.00	782.95	90.00	772.59	100.00
14	**fluid**	word	696.81	100.00	626.89	100.00	685.93	100.00	710.77	100.00
15	**fray**	word	908.48	91.67	857.49	88.89	890.58	75.00	1015.29	66.13
16	**hasty**	word	808.05	96.67	755.75	100.00	1006.08	70.00	852.70	100.00
17	**hurricane**	word	690.88	98.33	630.30	100.00	719.74	96.67	746.96	98.39
18	**ingenious**	word	884.40	100.00	746.77	97.22	698.08	100.00	1025.98	95.16
19	**lengthy**	word	699.83	98.33	686.17	100.00	804.77	95.00	1213.74	95.16
20	**listless**	word	937.16	93.33	874.92	83.33	1081.44	76.67	1338.70	72.58
21	**lofty**	word	896.02	90.00	878.69	83.33	1005.37	78.33	1164.10	93.55
22	**majestic**	word	777.69	100.00	667.91	100.00	693.82	98.33	794.90	96.77
23	**moonlit**	word	770.91	98.33	786.19	100.00	981.95	70.00	1078.71	77.42
24	**muddy**	word	681.71	100.00	644.40	100.00	822.40	83.33	790.64	98.39
25	**nourishment**	word	775.22	100.00	692.31	97.22	935.32	95.00	1048.37	96.77
26	**plaintively**	word	1216.37	76.67	1399.40	72.22	1202.08	83.33	1591.95	41.94
27	**rascal**	word	780.22	96.67	708.94	97.22	937.88	66.67	1018.35	75.81
28	**recipient**	word	766.49	100.00	737.16	100.00	771.78	100.00	874.34	96.77
29	**savory**	word	968.11	95.00	1170.06	91.67	879.07	85.00	795.35	93.55
30	**scholar**	word	726.26	98.33	766.12	100.00	753.22	93.33	828.39	96.77
31	**scornful**	word	784.60	98.33	883.92	97.22	1045.70	75.00	1175.45	91.94
32	**screech**	word	769.06	100.00	778.63	94.44	1068.41	70.00	1033.82	79.03
33	**shin**	word	810.76	96.67	795.52	97.22	974.89	68.33	947.16	77.42
34	**slain**	word	796.20	88.33	886.85	86.11	915.14	65.00	875.86	66.13
35	**stoutly**	word	1118.90	71.67	882.58	69.44	983.86	31.67	1235.60	56.45
36	**turmoil**	word	712.46	100.00	673.03	97.22	846.26	71.67	1105.83	85.48
37	**turtle**	word	691.12	100.00	734.93	100.00	693.51	100.00	778.65	100.00

			EN		EN (OX)		FR		KR	
	Stimulus	***Answer***	***Mean RT***	*%*	***Mean RT***	*%*	***Mean RT***	*%*	***Mean RT***	*%*
38	**unkempt**	word	918.42	80.00	901.90	94.44	994.69	41.67	1241.83	43.55
39	**upkeep**	word	794.60	98.33	768.82	97.22	963.41	68.33	1090.69	80.65
40	**wrought**	word	861.47	83.33	950.24	80.56	988.83	50.00	1240.82	50.00
41	abergy	non	881.63	0.00	930.72	11.11	884.73	8.33	1093.13	11.29
42	alberation	non	1251.69	21.67	1292.87	19.44	1152.71	10.00	1502.51	33.87
43	crumper	non	941.72	16.67	922.91	0	928.36	26.67	1079.67	25.81
44	destription	non	1141.12	16.67	1272.88	13.89	1076.68	31.67	1445.90	33.87
45	exprate	non	1073.07	10.00	1080.72	0	851.50	6.67	1300.08	17.74
46	fellick	non	872.17	3.33	875.44	0	783.78	5.00	1049.54	3.23
47	interfate	non	1235.24	11.67	1281.17	19.44	1059.69	28.33	1290.11	33.87
48	kermshaw	non	925.89	1.67	873.73	2.78	913.45	1.67	969.15	4.84
49	kilp	non	885.17	1.67	800.91	5.56	845.18	6.67	1145.83	3.23
50	magrity	non	924.08	3.33	885.15	5.56	909.36	5.00	1177.73	14.52
51	mensible	non	969.84	11.67	879.47	5.56	979.03	8.33	1397.48	20.97
52	plaudate	non	1062.67	8.33	1702.09	25.00	894.40	8.33	1370.78	14.52
53	proom	non	871.67	5.00	883.99	5.56	819.32	16.67	994.25	11.29
54	pudour	non	807.21	1.67	926.62	5.56	923.85	15.00	959.30	8.06
55	pulsh	non	988.40	8.33	917.92	2.78	929.98	11.67	1240.13	14.52
56	purrage	non	869.59	3.33	863.78	5.56	946.58	13.33	1141.61	11.29
57	quirty	non	1038.01	18.33	1167.38	5.56	964.05	23.33	1288.03	14.52
58	rebondicate	non	1266.71	1.67	1594.73	8.33	1082.17	23.33	1551.19	17.74
59	skave	non	928.86	8.33	933.82	5.56	799.35	3.33	970.66	11.29
60	spaunch	non	903.32	1.67	950.12	8.33	935.58	13.33	1193.80	9.68

Appendix B.4.2 French LexTALE

Table B.20: French LexTALE stimuli by word/non-word, with approximate English translations, average response times in milliseconds (ms), and number/percentage of selections.

Stimulus	Approximate Translation	Answer	Mean RT (ms)	Selections	Percentage
abêtir	(to) dull the mind; stupefy	word	934.58	26	43.33
alourdir	(to) burden, weigh down	word	829.24	57	95.00
amadouer	coax	word	799.88	58	96.67
amorce	initiation	word	800.41	54	90.00
balai	(a) broom	word	686.41	55	91.67
bouilloire	kettle	word	703.99	60	100.00
bouton	(a) button	word	641.64	58	96.67
caddie	caddy	word	721.06	56	93.33
cadenas	padlock	word	668.62	60	100.00
canoter	(to go) boating, rowing, or canoeing	word	1064.48	22	36.67
capeline	wide-brimmed hat	word	886.86	26	43.33
chameau	camel	word	721.87	59	98.33
cheveux	hair	word	589.18	60	100.00
cintre	hanger	word	742.57	56	93.33
citrouille	pumpkin	word	636.70	59	98.33
cloche	bell	word	668.36	60	100.00
clouer	(to) nail	word	840.21	59	98.33
crayon	pencil	word	639.45	60	100.00
dauphin	dolphin	word	622.97	60	100.00
écureuil	(a) squirrel	word	843.34	58	96.67

Stimulus	Approximate Translation	Answer	Mean RT (ms)	Selections	Percentage
église	church	word	633.91	60	100.00
ennemi	enemy	word	664.57	59	98.33
escroc	crook	word	668.79	60	100.00
esquif	skiff	word	931.05	22	36.67
éventail	(a) fan, OR: (a) range, variety	word	722.02	59	98.33
fascine	(a) bundle of sticks	word	901.60	47	78.33
fenêtre	window	word	609.32	60	100.00
fosse	(a) pit	word	726.81	58	96.67
fouet	(a) whip	word	692.38	59	98.33
fourmi	ant	word	833.57	59	98.33
hache	(an) axe	word	643.62	57	95.00
inciter	encourage	word	750.02	58	96.67
indicible	unspeakable	word	979.48	42	70.00
infâme	infamous	word	754.38	57	95.00
lanière	(a) strap, thong, lash	word	770.55	57	95.00
lézard	lizard	word	688.69	59	98.33
mappemonde	world map	word	954.61	43	71.67
marteau	(a) hammer	word	667.72	60	100.00
mignon	cute	word	651.32	60	100.00
nouer	(to) tie	word	720.24	60	100.00
occire	slay	word	877.85	37	61.67
oeillet	carnation	word	698.50	57	95.00
orgueil	pride	word	680.82	60	100.00
panier	basket	word	639.33	60	100.00
peigne	(a) comb	word	656.26	60	100.00
pinceau	(a) brush	word	702.67	60	100.00
poisson	(a) fish	word	595.73	59	98.33
pouce	(an) inch	word	628.69	60	100.00
racaille	riffraff	word	733.73	59	98.33

Stimulus	Approximate Translation	Answer	Mean RT (ms)	Selections	Percentage
remporter	(to) win	word	724.37	60	100.00
robinet	(a) tap / faucet	word	726.71	58	96.67
salière	saltshaker	word	927.56	57	95.00
semonce	(a) reprimand, reproach	word	960.72	29	48.33
tanin	tannin	word	888.16	41	68.33
treillage	trellis	word	1114.93	35	58.33
vicelard	(a) pervert, lecher; lecherous	word	880.08	53	88.33
agire	-	non-word	976.48	17	28.33
boutard	-	non-word	1012.00	10	16.67
cerveler	-	non-word	1098.78	16	26.67
cessure	-	non-word	1048.08	25	41.67
détume	-	non-word	791.21	1	1.67
écouce	-	non-word	832.47	4	6.67
endifier	-	non-word	918.11	2	3.33
gloque	-	non-word	1071.59	19	31.67
honteur	-	non-word	907.41	16	26.67
huif	-	non-word	905.32	10	16.67
jamain	-	non-word	795.63	1	1.67
joueux	-	non-word	813.91	22	36.67
metter	-	non-word	846.46	8	13.33
oeiller	-	non-word	1093.30	22	36.67
osseaux	-	non-word	1015.84	12	20.00
parchance	-	non-word	971.12	12	20.00
parir	-	non-word	897.98	5	8.33
plaiser	-	non-word	920.28	11	18.33
pourcine	-	non-word	941.47	9	15.00
prioche	-	non-word	871.88	3	5.00
procoreux	-	non-word	865.10	2	3.33
rejoute	-	non-word	927.09	8	13.33
replaner	-	non-word	1123.14	15	25.00

Stimulus	Approximate Translation	Answer	Mean RT (ms)	Selections	Percentage
réporce	-	non-word	817.36	3	5.00
retruire	-	non-word	873.88	7	11.67
sacher	-	non-word	912.97	13	21.67
sentuelle	-	non-word	892.40	7	11.67
soumon	-	non-word	800.49	6	10.00

Appendix B.4.3 Korean LexTALE

Table B.21: Korean LexTALE stimuli by word/non-word, with approximate English translations, average response times in milliseconds (ms), and number/percentage of selections.

Stimulus	Approx. Translation	Answer	Mean RT (ms)	Selections	Percentage
가운데	middle, centre	word	664.10	61	98.39
개발	development	word	704.80	57	91.94
결국	eventually, finally	word	686.23	62	100.00
과학	science	word	693.11	61	98.39
관련	related, relevant	word	715.46	58	93.55
교육	education	word	638.61	61	98.39
굶다	to starve	word	759.27	59	95.16
권리	right (as in human rights)	word	750.88	58	93.55
기쁨	joy	word	636.35	61	98.39
까닭	reason, cause	word	1030.95	52	83.87
꿰차다	to occupy, to take up	word	1083.80	49	79.03
나오다	to come out, to appear	word	649.35	61	98.39
논란	controversy	word	664.30	60	96.77

Stimulus	Approx. Translation	Answer	Mean RT (ms)	Selections	Percentage
누더기	rag, tatter	word	853.85	53	85.48
다짐	promise, resolution	word	677.15	58	93.55
닿다	to touch, to reach	word	900.20	51	82.26
등극	ascension, enthronement	word	886.36	40	64.52
또래	peer, same age group	word	915.98	59	95.16
머금다	to hold (in the mouth), to contain	word	926.88	53	85.48
모르다	to not know	word	673.38	61	98.39
민족	ethnic group, nation	word	650.92	58	93.55
바깥	outside	word	788.05	57	91.94
변화	change	word	670.53	60	96.77
부르다	to call, to sing	word	699.95	61	98.39
분명	clear, obvious	word	634.24	61	98.39
사람	person	word	637.15	61	98.39
살다	to live	word	681.87	62	100.00
새롭다	new	word	738.66	61	98.39
생기다	to happen, to look like	word	685.82	62	100.00
숱하다	numerous, many	word	977.73	43	69.35
아늑히	cosy, snug	word	936.62	46	74.19
알다	to know	word	733.59	62	100.00
얼굴	face	word	648.16	62	100.00
없다	to not exist, to not have	word	731.78	61	98.39
엉터리	nonsense, rubbish	word	753.62	59	95.16

Stimulus	Approx. Translation	Answer	Mean RT (ms)	Selections	Percentage
여백	margin, blank space	word	751.81	53	85.48
왜곡	distortion	word	824.37	51	82.26
원칙	principle	word	698.85	60	96.77
잃다	to lose	word	699.78	60	96.77
정착	settlement	word	747.26	59	95.16
젖니	milk tooth	word	1171.25	35	56.45
제각각	each, respectively	word	1049.89	47	75.81
존재	existence	word	660.60	61	98.39
주다	to give	word	741.92	61	98.39
중략	omission (in text)	word	860.84	53	85.48
찾다	to find, to look for	word	664.06	62	100.00
출몰	appearance, frequenting	word	863.99	51	82.26
특히	especially	word	707.97	62	100.00
파괴	destruction	word	714.52	56	90.32
판매	sale	word	717.98	60	96.77
표현	expression	word	691.50	59	95.16
필요	necessity	word	661.16	62	100.00
함께	together	word	631.60	60	96.77
현미경	microscope	word	817.33	57	91.94
혐의	suspicion	word	767.97	55	88.71
회담	conference, talk	word	748.35	57	91.94
흡사	similar	word	721.13	55	88.71
같티	-	non-word	932.28	3	4.84
거늬	-	non-word	756.56	1	1.61
곤계	-	non-word	825.19	6	9.68
과더	-	non-word	799.08	3	4.84

Stimulus	Approx. Translation	Answer	Mean RT (ms)	Selections	Percentage
국밀	-	non-word	1085.00	3	4.84
규노	-	non-word	850.91	1	1.61
기촌	-	non-word	930.36	9	14.52
꼬링	-	non-word	755.22	1	1.61
남히	-	non-word	904.07	8	12.90
노알	-	non-word	872.18	2	3.23
다늘	-	non-word	917.64	3	4.84
다지막	-	non-word	840.29	7	11.29
다프다	-	non-word	829.78	1	1.61
단붕	-	non-word	825.51	5	8.06
단지다	-	non-word	920.08	9	14.52
디로	-	non-word	875.77	2	3.23
마츰	-	non-word	978.45	5	8.06
모징	-	non-word	900.99	0	0.00
목조리	-	non-word	1088.11	6	9.68
문위기	-	non-word	988.80	4	6.45
물곤	-	non-word	1044.00	2	3.23
뭅다	-	non-word	828.49	6	9.68
바남	-	non-word	926.15	0	0.00
별노	-	non-word	1104.35	1	1.61
부히	-	non-word	901.65	3	4.84
빌통	-	non-word	887.01	4	6.45
소묵	-	non-word	915.85	7	11.29
신덩서	-	non-word	826.06	1	1.61
씬다	-	non-word	896.42	6	9.68
예숙	-	non-word	1039.37	8	12.90
완벙	-	non-word	797.02	3	4.84
잔끅	-	non-word	812.36	7	11.29
잭다	-	non-word	814.11	0	0.00

Stimulus	Approx. Translation	Answer	Mean RT (ms)	Selections	Percentage
종됴	-	non-word	778.83	1	1.61
창다	-	non-word	910.87	4	6.45
통억	-	non-word	846.24	5	8.06
포잔	-	non-word	934.43	2	3.23
하매다	-	non-word	964.18	2	3.23
호떨	-	non-word	770.62	4	6.45
횡병	-	non-word	1109.43	6	9.68
후민	-	non-word	792.44	7	11.29

Appendix B.5 Collocations (Chapters 3 & 4)

Table B.22: Item characteristics for the collocations task, with percentage of correct selections (%) by cohort.

Note: Frequency information, z-scores, t-scores, and mutual information (MI) values are from Dąbrowska (2014), and were extracted from the British National Corpus.

Item	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	*z*-score	*t*-score	MI	% EN	% EN (OX)	% FR	% KR
absolute silence	27	3377	5162	0.18	63.3	5.2	7.2	73.33	61.11	43.33	58.06
achieve one's objectives	316	16553	7210	1.23	283.8	17.7	8	76.67	52.78	66.67	54.84
arouse suspicions	73	1330	2130	0.03	427	8.5	11.3	96.67	97.22	63.33	74.19
attract publicity	38	6229	2442	0.16	95.6	6.1	7.9	90.00	80.56	43.33	41.94
attractive proposition	48	5000	1987	0.10	149.7	6.9	8.9	68.33	38.89	35.00	38.71
bend rules	27	3259	18225	0.61	33.7	5.1	5.5	93.33	86.11	60.00	51.61
bitter dispute	54	2357	4435	0.11	164.2	7.3	9	88.33	44.44	11.67	43.55

Item	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	*z*-score	*t*-score	MI	% EN	% EN (OX)	% FR	% KR
blank expression	13	1393	8456	0.12	37	3.6	6.7	93.33	88.89	46.67	56.45
blatant lie	12	323	2150	0.01	141.7	3.5	10.7	98.33	100	61.67	59.68
boost production	93	1676	15791	0.27	177.5	9.6	8.4	90.00	94.44	68.33	64.52
close similarity	15	10326	1667	0.18	35.2	3.8	6.4	65.00	30.56	38.33	40.32
dim view	47	678	28193	0.20	105.4	6.8	7.9	65.00	66.67	20.00	16.13
divert attention	154	1156	13457	0.16	384.1	12.4	9.9	70.00	61.11	20.00	46.77
divert suspicion	6	1156	2130	0.03	37.5	2.4	7.9	76.67	72.22	35.00	37.10
fair share	272	7870	15830	1.28	238.9	16.4	7.7	98.33	97.22	80.00	83.87
full confession	10	27288	832	0.23	20.2	3.1	5.4	80.00	80.56	70.00	74.19
gain popularity	43	8601	1304	0.12	126.1	6.5	8.5	83.33	88.89	71.67	66.13

Item	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	*z*-score	*t*-score	MI	% EN	% EN (OX)	% FR	% KR
general direction	93	29308	10505	3.17	50.4	9.3	4.9	86.67	83.33	30.00	41.94
hazard a guess	44	110	799	0	1461.6	6.6	15.6	85.00	94.44	28.33	20.97
hear rumours	99	34199	1853	0.65	121.7	9.9	7.2	73.33	80.56	70.00	58.06
inflict punishment	18	1027	2423	0.03	112.2	4.2	9.5	65.00	66.67	55.00	45.16
issue a statement	390	7833	13648	1.10	370.4	19.7	8.5	86.67	86.11	63.33	54.84
join the ranks	97	16701	3433	0.59	125.4	9.8	7.4	76.67	66.67	76.67	61.29
lodge a complaint	33	1066	4425	0.05	149.4	5.7	9.4	81.67	69.44	31.67	30.65
memorable phrase	13	832	4143	0.04	68.8	3.6	8.5	70.00	50.00	25.00	46.77
obvious conclusion	36	8234	7320	0.62	44.9	5.9	5.9	51.67	36.11	61.67	37.10
odd remark	8	4255	3049	0.13	21.5	2.8	5.9	61.67	66.67	41.67	59.68

Item	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	*z*-score	*t*-score	MI	% EN	% EN (OX)	% FR	% KR
outspoken critic	41	293	3690	0.01	388.2	6.4	11.8	53.33	36.11	31.67	29.03
overall responsibility	93	5897	11809	0.72	108.9	9.6	7	63.33	22.22	41.67	33.87
precise details	67	2834	17294	0.51	93.5	8.1	7.1	55.00	61.11	35.00	35.48
raise prices	109	18786	27440	5.32	45	9.9	4.4	91.67	94.44	75.00	88.71
raise standards	173	18786	14878	2.88	100.2	12.9	5.9	66.67	88.89	43.33	58.06
refuse an application	82	10172	15869	1.66	62.3	8.9	5.6	36.67	30.56	6.67	35.48
regular employment	31	7387	10600	0.81	33.6	5.4	5.3	86.67	80.56	75.00	80.65
restore faith	25	3839	5160	0.20	54.9	5	6.9	60.00	41.67	46.67	43.55
serious problem	619	11903	54555	6.70	236.6	24.6	6.5	68.33	77.78	56.67	70.97
striking example	92	1667	19265	0.33	159.3	9.6	8.1	41.67	47.22	43.33	22.58

Item	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	*z*-score	*t*-score	MI	% EN	% EN (OX)	% FR	% KR
thorough search	25	1081	5378	0.06	101.9	5	8.7	75.00	83.33	45.00	54.84
urgent matters	36	2066	23720	0.51	49.9	5.9	6.2	68.33	77.78	38.33	43.55
witness an incident	20	2015	5033	0.1	61.5	4.4	7.6	81.67	86.11	78.33	64.52

End Matter

Biographical Note

Sean Patrick McCarron was born in London, Ontario, Canada, and developed an interest in books and language from an early age, with a particular love for comic books and fantasy. This sparked an enduring interest in creative language use and non-standard dialects. Over the years he has worn many hats (sometimes literally), working for years as a contract illustrator, drawing storyboards, film production artwork, colouring books, and logo designs, in addition to self-publishing a series of all-ages graphic novels. Becoming bilingual in French in late adolescence inspired him to better understand how the brain learns and processes language, leading to an Honours B.A. and M.Sc. in Cognitive Science of Language at McMaster University in Hamilton, Ontario, where he met his wife, Lindsay. While living in Oxford, UK for three years during the writing of this thesis, he and Lindsay raised their four young children, who remain by far his best work.