The Potential Influence of Crosslinguistic Similarity on Lexical Transfer: Examining Vocabulary Use in L2 English

By Itamar Shatz

UNIVERSITY OF CAMBRIDGE

(HUGHES HALL)

This thesis is submitted for the degree of *Doctor of Philosophy*

November, 2021

ABSTRACT

Learners' native language (L1) influences their knowledge and use of second language (L2) vocabulary, a phenomenon known as *lexical transfer*. Past research on this shows that learners' L1 influences their L2 word choices, and that lexical similarity—which relates to cognancy—between L1 words and their L2 counterparts facilitates the processing of the L2 words, particularly during the early stages of L2 acquisition, and makes speakers more likely to use the L2 words in spontaneous productions.

To extend past research, the present research investigates whether crosslinguistic similarity influences L2 vocabulary use in a task-based, English-as-a-foreign language educational setting. Specifically, it investigates whether increased similarity between languages as a whole increases L2 lexical diversity, and whether increased similarity between L1 words and their L2 counterparts increases the use of the L2 words. It investigates this using two matching learner samples, containing 8,500 and 6,390 English texts, written in response to 95 and 71 tasks, by speakers of 9 typologically diverse L1s, in the A1–B2 CEFR range of L2 proficiency.

Surprisingly, lexical similarity between the L1 and the L2 as a whole did not influence L2 lexical diversity, regardless of learners' L2 proficiency. Likewise, lexical similarity between corresponding L1-L2 words did not influence the use of the L2 words, again regardless of L2 proficiency. Conversely, there were strong task effects on both L2 lexical diversity and L2 word choice.

These findings show that the facilitative effect of crosslinguistic lexical similarity (especially the cognate facilitation effect) is constrained, and suggest that communicative needs and other task effects can override positive lexical transfer. This highlights the role of situational factors in crosslinguistic influence, and raises questions regarding when and how these and similar factors can override language transfer, for example when it comes to different types of transfer (e.g., positive vs. negative, or lexical vs. syntactic). In addition, this research contains substantial insights into related topics, such as the developmental patterns of L2 lexical diversity, accounting for task effects in language assessment, measuring crosslinguistic distance, and using online platforms to develop language corpora.

TABLE OF CONTENTS

DECLARATION

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. No substantial part of it has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dora Alexopoulou, for all her help and guidance throughout the PhD. Likewise, I would like to thank Akira Murakami, for all his valuable comments and insights. Many thanks also go to everyone else who provided helpful feedback and discussions, in the course of things such as conference presentations and journal reviews.

This research used various open-access datasets, software, and guides; I would like to sincerely thank everyone involved in creating these sources and making them accessible. On a similar note, I would like to thank everyone who provided resource support for this research, including EF-Education First, Cambridge Assessment English, Hughes Hall, and the University of Cambridge.

On a more personal note, I would like to thank my family, friends, and everyone else who was there, with a special thanks to my main training partner.

Finally:

Mom and dad—Thank you for everything. I love you both.

My amazing wife—It has been quite a journey, and I look forward to everything that comes next. I love you so much.

1 Introduction

1.1 Crosslinguistic influence and language transfer

Crosslinguistic influence (or *language transfer*) occurs when people's knowledge of one language influences their knowledge and use of another language.[1] A distinction is often made between *positive transfer*, which facilitates learners' engagement with a language in terms of factors such as processing, acquisition, and use, and *negative transfer* (or *interference*), which hinders learners' engagement with a language (Benson, 2002; Cebrian, 2000; R. Ellis, 2008; Gujord, 2020; L. S. Huang, 2009; James, 2012; Jarvis, 2000, 2009, 2015, 2017; Jarvis & Pavlenko, 2008; Kellerman, 1995; Koda, 2005; Kubota, 1998; Llach, 2010; O'Sullivan & Chambers, 2006; Odlin, 1989, 2003, 2005, 2013; Ringbom, 1987, 1992, 2007; Sersen, 2011; Yuan, 2014).

Crosslinguistic influence can occur between any combination of languages, but most research focuses on influence from people's native language (L1) to another language that they know (L2) (Jarvis, 2017).[2] This influence can occur in all linguistic domains, such as phonology, semantics, morphology, and syntax. Various factors, such as the linguistic domain involved and crosslinguistic similarities, can determine whether learners experience influence, and if so then in what way and to what degree.[3] For example, in the domain of morphosyntax, grammatical features such as the neuter gender are acquired by learners faster when they are instantiated in a similar manner in learners' L1 and their target L2, compared to when they are instantiated differently, or when they appear only in the L2 (Tolentino & Tokowicz, 2011, 2014).

In the context of L2 acquisition, crosslinguistic influence is often said to affect learners' *interlanguage*—their developing L2 knowledge, including their L2 grammar and lexicon—which depends on factors such as their target L2, their L1, and any additional languages that they speak (Cebrian, 2000; Corder, 1967; Gudmestad, 2012; Gujord, 2020; Han & Tarone, 2014; Ipek, 2009; Jarvis & Pavlenko, 2008; Kumpf, 1984; Lim, 2007; Llach, 2010; Major,

[1] Though this phenomenon does not necessarily involve transfer of linguistic features or patterns across languages, these two terms are generally used interchangeably, primarily because "transfer" can be more convenient to use and has become entrenched in the literature (Jarvis, 2017; Odlin, 2013).

[2] In line with much of the literature, I use "L2" here to refer to any language that is acquired after a learner's L1, though it may not necessarily be the second language that the learner has acquired (Jarvis, 2017).

[3] In saying that learners "experience" crosslinguistic influence, I do not suggest that this is something that they are necessarily conscious of, as learners are often unaware of this influence (Benson, 2002; Llach, 2010; Ringbom, 2007). Nevertheless, being consciously awareness of one's language use and of related concepts, such as crosslinguistic similarities, can affect the crosslinguistic influence that people experience (Benson, 2002; Jarvis & Pavlenko, 2008; N. Jiang, 2002).

1998; Murphy, 2003; Odlin, 2003; Selinker, 1972, 2013). Crosslinguistic influence was originally believed to affect learners' interlanguage mostly during the initial stages of acquisition, but later research showed that L1 influence can also play a substantial role during advanced stages of L2 acquisition (R. Ellis, 2008; Jarvis & Pavlenko, 2008; Llach, 2010; Ringbom, 2007; Schwartz & Sprouse, 1996; Upton & Lee-Thompson, 1987).

In summary, learners' L1 can influence the way in which they acquire, process, and use an L2, a phenomenon known as *crosslinguistic influence* or *language transfer*. This occurs primarily—but not exclusively—during the initial stages of acquisition, when learners' L2 proficiency is low. The nature of this influence is highly variable, and depends on many factors beyond learners' L2 proficiency, such as the linguistic domain under consideration, and the similarities between learners' L1 and their target L2.

1.2 Lexical transfer

[This section provides an overview of lexical transfer in the context of the present research. For more information on lexical transfer, including additional examples and a discussion of some relevant theories, see Appendix A.]

Lexical transfer is a type of crosslinguistic influence that occurs when people's knowledge of one language influences their knowledge and engagement with another language's vocabulary, in terms of operations such as recognition, interpretation, storage, and retrieval (Jarvis, 2009).

A notable aspect of lexical transfer is that crosslinguistic similarity in form between L1 words and their L2 translations—which are referred to as psycholinguistic *cognates* when the similarity in both meaning and form is high—facilitates the processing and learning of the L2 words (Bosma et al., 2019; Costa et al., 2000, 2005; de Groot & Keijzer, 2000; C. J. Hall, 2002; Helms-Park & Dronjic, 2013; Jarvis, 2009; N. Jiang, 2002; Lindgren & Bohnacker, 2020; Llach, 2010; Lotto & de Groot, 1998; Poort & Rodd, 2017; Ringbom, 2007; Sánchez-Casas & García-Albea, 2005; Tonzar et al., 2009; Vanlangendonck et al., 2020; Williams, 2015; Y. Zhu & Mok, 2020).[4] This *cognate facilitation effect* means that when a word in the learner's L1 has

[4] This interpretation of the term *cognate* is most common in fields such as psycholinguistics and language acquisition, which are primarily concerned with crosslinguistic similarities and differences as they are perceived by learners. However, the term *cognate* can also refer to words in different languages that share similar etymological origins; this meaning is most common in fields that are interested in such relations, such as historical linguistics and linguistic typology. Nevertheless, etymological cognancy often leads to psycholinguistic cognancy (Carrasco-Ortiz et al., 2021; Rabinovich et al., 2018; Schepens et al., 2012), so there is often—but not always—an overlap between these two forms of cognancy, meaning that words that are cognate in one of these senses are also cognate in the other.

phonological and/or orthographic similarity to its L2 counterpart, the learner will generally have an easier time processing, acquiring, and using the L2 word. For example, for a French speaker, it will likely be easier to learn the English word *orange*, for which the corresponding word in French is also *orange*, than to learn the word *lemon*, for which the corresponding word in French is *citron*.

The cognate facilitation effect, which is well-attested in the psycholinguistic and second language acquisition (SLA) literature, has been attributed to various cognitive mechanisms. However, the general explanation for it is that similarity in form between L1 words and L2 words that share their meaning facilitates the linking and/or mapping of L2 words to their L1 counterparts or to shared concepts, which facilitates the transfer of linguistic (e.g., semantic, syntactic, and morphological) information from the L1 to the L2 (Costa et al., 2000; Ecke, 2015; C. J. Hall, 2002; Helms-Park & Dronjic, 2013; Jarvis & Pavlenko, 2008; N. Jiang, 2002; Ringbom, 2007; Tonzar et al., 2009). Furthermore, most approaches suggest that this type of lexical transfer plays a role primarily—but not exclusively—during early stages of L2 acquisition (C. J. Hall, 2002; Jarvis, 2009; Llach, 2010; Williams, 2015). Overall, Ringbom (2007) summarizes this phenomenon as follows:

> Formal similarities, phonological and orthographical, have an essential role in the organisation of the mental lexicon, especially at early stages of learning. These similarities may be predominantly cross-linguistic or predominantly intralinguistic, with the proportion being determined largely by the distance perceived between L1 and L2 and by the proficiency of the learner. (p. 28)

However, though there is much research on the topic, most of it has focused on experimental assessment of L2 processing, and on a limited range of L1s and L2 proficiency levels (Rabinovich et al., 2018). For example:

- C. J. Hall (2002) examined 95 Spanish-speaking university students in an intermediate English for Academic Purposes course, and found that they reported higher levels of familiarity with pseudo-cognates than non-cognates, and showed greater consistency in translating the pseudo-cognates.
- N. Jiang (2002) examined 25 Chinese-English bilingual speakers who were graduate students at an American university, and found that they responded faster than native English speakers to L2 word pairs that share the same L1 translations than to pairs that do not.

– Tonzar et al. (2009) examined 229 Italian speaking children, and found that they processed L2 cognates more accurately than non-cognates.

Some research did examine the influence of crosslinguistic lexical similarity using broad samples, and found that increased similarity between languages as a whole—based on the mean similarity between corresponding L1-L2 words—leads to higher overall L2 proficiency scores in both speaking (Schepens et al., 2020; Schepens, van der Slik, et al., 2013b) and writing (van der Slik, 2010). This could be due to facilitated processing and/or production of L2 words, for example if the facilitated ability to process certain L2 words makes it easier for learners to acquire these words, and also frees up cognitive resources that can then be used to acquire other L2 words, which aligns with concepts such as comprehensible input (Krashen, 1989, 2003).[5] However, because research on this examined composite L2 scores that include a mix of factors such as vocabulary, grammar, and pronunciation, it is unclear what role lexical transfer plays here, and particularly whether it influences learners' L2 productions directly.

In addition, some research did look at L2 productions, and found evidence of a specific type of lexical transfer, called *word choice transfer*, whereby a person's knowledge of a language influences their choice of words in another language (Jarvis et al., 2012; Jarvis & Pavlenko, 2008; Kyle et al., 2015; Stemle & Onysko, 2015). Based on this research, learners' use of specific words and phrases, which is referred to as *lexical signature*, *lexical style*, or *wordprints*, has been shown to be relevant for use in stylometry for L1 identification, in both spontaneous productions and task-based productions such as TOEFL essays (Jarvis et al., 2012; Jarvis & Pavlenko, 2008). However, research on this did not generally investigate whether this phenomenon is driven by crosslinguistic similarity in particular, or by other factors, such as a cultural preference for certain words.

The key exception that sounds out in this regard is Rabinovich et al. (2018), who investigated the influence of L1-L2 similarity on L2 word choice. They did so using a relatively

[5] Some support for this comes from Ard and Homburg (1983), who found that lexical similarity between learners' L1 and the target L2 can facilitate the acquisition of vocabulary items that are not a part of an L1-L2 cognate pair. This is based on the observation that Spanish learners of English as a second language (ESL) were better able to learn English words than Arabic ESL learners, even when it comes to L2 words that are not similar to their L1 counterparts. They attribute this to the increased similarity between Spanish and English compared to Arabic and English, and propose two mechanisms through which such similarity could facilitate the acquisition of the non-similar words. First, it is possible that because learners have an easier time learning words that are similar between the two languages, they can dedicate more time and mental resources to learning words that are not similar, which leads to improved rates of acquisition for those types of words. Second, it is possible that even when words are not directly similar between the L1 and the L2, languages that have more similar lexical items tend to also have more similar lexical structures, in terms of their phonology and orthography, which could facilitate acquisition of L2 words, even if they're not a part of a cognate pair.

large and broad sample, involving the spontaneous productions on social media of highly proficient (near-native) L2 English speakers, who represented a wide range of Indo-European L1s. The researchers found clear evidence of the cognate facilitation effect, since speakers were more likely to use L2 words when they were cognate with their L1 counterparts. However, this research still contains some gaps, for example when it comes to investigating how L2 proficiency and task effects can moderate this cognate facilitation effect.

Investigating the roles of task is particularly important, since these findings contrast with those of Crossley and McNamara (2011), who found no L1 effect when it comes to several global L2 lexical measures, such as lexical diversity and polysemy. This is based on an examination of 599 L2 English texts in the *International Corpus of Learner English*, written by Czech, Finnish, German, and Spanish L1 speakers with high-intermediate to advanced L2 English proficiency in response to one of few prompts for argumentative essays. This finding of *intergroup homogeneity* is unexpected given the L1 effects that were found in other studies, and casts uncertainty regarding the extent to which learners' L1 influences learners' L2 lexical productions. However, this finding does not necessarily contradict other past findings, since it is the only one to examine L1 effects on this type of global lexical measure in a task-based setting. Furthermore, while the lack of L1 effect here suggests that L1-L2 similarity might not influence these global lexical measures, at least in the task-based setting examined by these researchers, the researchers did not actually consider crosslinguistic similarity in their analyses, so there is uncertainty regarding whether the L1s are sufficiently different in their lexical similarity from English to prompt different levels of similarity-based L2 facilitation.

In summary, past research on lexical transfer indicate the following key things, in relation to the present research:

- Similarity in form (i.e., phonology and/or orthography) between L1 words and their L2 counterparts facilitates the processing of the L2 words, particularly—but not exclusively—at early stages of L2 acquisition (Bosma et al., 2019; Costa et al., 2000, 2005; de Groot & Keijzer, 2000; C. J. Hall, 2002; Helms-Park & Dronjic, 2013; Jarvis, 2009; N. Jiang, 2002; Lindgren & Bohnacker, 2020; Llach, 2010; Lotto & de Groot, 1998; Poort & Rodd, 2017; Ringbom, 2007; Sánchez-Casas & García-Albea, 2005; Tonzar et al., 2009; Vanlangendonck et al., 2020; Williams, 2015; Y. Zhu & Mok, 2020).
- Increased L1-L2 lexical similarity between languages as a whole, based on the mean similarity of corresponding words, leads to higher overall L2 proficiency scores, but it

is not clear what role lexical transfer plays in this (Schepens et al., 2020; Schepens, van der Slik, et al., 2013b; van der Slik, 2010).

- Learners' L1 can influence their choice of L2 words in both spontaneous and task-based settings, though it is unclear whether this effect is driven by crosslinguistic lexical similarity, and if so then to what degree (Jarvis et al., 2012; Jarvis & Pavlenko, 2008; Kyle et al., 2015; Stemle & Onysko, 2015).
- Highly proficient L2 speakers are more likely to use L2 words that are cognate with their L1 in spontaneous productions, though there is limited research on this (Rabinovich et al., 2018).
- Learners' L1 does not influence their lexical diversity in task-based settings, though there is limited research on this (Crossley & McNamara, 2011).

1.3 Research questions

Overall, based on prior research, it is clear that learners' L1 can influence their L2 in various ways. Most notably, in the context of lexical transfer, there is a robust facilitative effect of L1-L2 lexical similarity on L2 processing and learning, as well as a robust effect of L1 on L2 word choice (with no claim regarding the influence of similarity in this regard). What is currently unclear, however, is whether there is an effect of L1-L2 lexical similarity on L2 vocabulary use in task-based settings, when it comes both to the choice of individual L2 words and to global lexical measures, such as lexical diversity.

The present research will address this gap in knowledge, by examining the influence of crosslinguistic similarity on lexical transfer in a task-based setting. Specifically, it will answer the following key research questions:

1. Does L1-L2 lexical similarity between languages as a whole influence L2 lexical diversity (in a task-based educational setting)?
2. Does L1-L2 similarity between L1 words and their L2 counterparts influence the use of the L2 words (in a task-based educational setting)?

The research will also examine other aspects of lexical transfer pertaining to these questions, such as whether such transfer, if it exists, is moderated by learners' L2 proficiency. In addition, the research will investigate other important aspects of L2 vocabulary use, such as the developmental patterns of L2 lexical diversity, and will shed light on important methodological concepts, such as the development of learner corpora from online platforms.

In doing all this, this research will benefit from the use of broad analyses, the need for which has been identified in prior research (Crossley & McNamara, 2011; Rabinovich et al., 2018). Specifically, this will involve the use of a broad sample, in terms of factors such as the number of learners, number of texts, number and type of tasks, number and diversity of L1s, and range of L2 proficiency levels involved. This will also involve the use of broad models, which control for a large range of relevant variables, such as L2 proficiency, crosslinguistic similarity, and task effects.

1.4 Thesis outline

This introduction is followed by three key chapters, each written in the form of a self-contained paper:

- In the first chapter, I present my learner sample and explain how I developed it, in a way that can inform the work of others who are developing and using similar corpora.
- In the second chapter, I investigate the influence of lexical similarity between languages as a whole on L2 lexical diversity (thus addressing my first research question).
- In the third chapter, I investigate the influence of the similarity between L1 words and their L2 counterparts on the use of the L2 words (thus addressing my second research question).

This is followed by a final chapter that discusses the key findings and implications of the two studies. This chapter also presents extensive suggestions for future research, including novel insights into many related topics, such as accounting for phonological weights and diacritics in calculations of lexical distance, and assessing the influence of segmental frequency and permissibility on lexical transfer.

In terms of structure, note that the “References” sections of the individual have been merged to a single section at the end of the thesis, to improve the flow of reading and avoid duplication. In addition, the two main studies make references to several supplementary documents; all these documents are included in the thesis as appendices, and the relevant appendix corresponding to each document is referenced in the first mention of the document.

2 Getting learner data: The EFCAMDAT Cleaned Subcorpus

This chapter of the thesis describes how I developed the learner sample that I used in my research, by modifying the *EF-Cambridge Open Language Database* (EFCAMDAT) to create the *EFCAMDAT Cleaned Subcorpus*.

Originally, I expected to work with the EFCAMDAT itself directly. However, when I looked at the dataset, I noticed some issues with it that I wanted to address before conducting my analyses. The quest to address those issues, which involved extensive organization and cleanup of the original dataset, ended up being substantial enough to be considered a research contribution in its own right, particularly as it can inform the work of others. As such, I published this chapter as the following paper, which discusses how I processed the EFCAMDAT, and is also framed as part of a broader discussion on the development and use of language corpora:

> Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 221–237.

In addition, I have made the resulting dataset, together with other relevant material and code, openly available on the EFCAMDAT website (https://corpus.mml.cam.ac.uk/), for use by other researchers, as noted in the paper.

The abstract for the paper is as follows, with the number of tokens added in brackets:

> This report outlines the development of a new corpus, which was created by refining and modifying the largest open-access L2 English learner database—the EFCAMDAT. The extensive data-curation process, which can inform the development and use of other corpora, included procedures such as converting the database from XML to a tabular format, and removing problematic markup tags and non-English texts. The final dataset contains two corresponding samples, written by similar learners in response to different prompts, which represents a unique research opportunity when it comes to analyzing task effects and conducting replication studies. Overall, the resulting corpus contains ~406,000 texts [~24,826,000 tokens] in the first sample and ~317,000 texts [~20,564,000 tokens] in the second sample, written by learners representing diverse L1s and a large range of L2 proficiency levels.

2.1 Introduction

Learner corpora are increasingly being developed from data that originates in large-scale online platforms. This is beneficial, since the growing size of such corpora enables the analysis of large amounts of learner data, in ways that were not possible before (Callies, 2015; McEnery et al., 2019). However, with these new data sources come new challenges, which require new developments in terms of how researchers curate and analyze learner corpora. One notable challenge is the need to develop approaches to working with data that was originally collected with educational or social goals in mind, rather than research, since such data is often messy and requires substantial processing before it can be properly analyzed. In addition, because of the increasing size of these new corpora, new approaches to data curation and analysis must be scalable, so they can be applied effectively on large-scale datasets, which puts an emphasis on the use of quantitative and NLP-based approaches.

The present report addresses this topic by discussing the development of a new derivative corpus from an existing learner language database. The goals of this report are both to introduce the new derivative source, and to explain how it was developed, in order to inform future data curation and analysis by researchers working with other learner corpora, and potentially with other corpora in general.

In particular, the original database used in this report is the *EF Cambridge Open Language Database* (EFCAMDAT), which is the largest open-access L2 English learner database, with 1,180,310 texts written by 174,743 learners from various nationalities (Geertzen et al., 2013; Y. Huang et al., 2017, 2018). The texts in the EFCAMDAT were submitted by learners to EF's online English school, which spans 16 English proficiency levels aligned with common proficiency standards such as the CEFR. Each level consists of 8 units, and upon completing a unit, learners are tasked with writing a text, which is then graded. If the learner receives a passing grade, they advance to the next unit; otherwise, they repeat the unit. The texts cover a variety of topics, such as reviewing a song for a website or describing one's favorite day.

The EFCAMDAT is pseudo-longitudinal overall, as learners generally complete only parts of the learning program. However, it contains substantial longitudinal data, since many learners complete sequences of tasks across increasing levels of proficiency, and researchers can track individual learners using the *learner ID* variable. In terms of metadata, the EFCAMDAT lists learners' English proficiency and their nationality, and learners were only added to the database if their nationality matched their country of residence (Alexopoulou et al., 2017). Prior research on the EFCAMDAT used learners' nationality to estimate their L1, an approach that has been validated empirically (Alexopoulou et al., 2017; Y. Huang et al., 2018; Murakami, 2014).

I developed the derivative version of the EFCAMDAT because I wanted to conduct a large-scale quantitative study of L2 lexical development, and found that I first needed to make several substantial modifications to the EFCAMDAT. As such, some of the decisions made in the course of creating the derivative corpus may not work well for other types of research. For example, the removal of duplicate texts described below may interfere with analyses that focus on formulaic language. However, researchers can choose to implement only some of the procedures that I outline in this report; to facilitate this, I make the relevant programmatic scripts available, together with partially cleaned versions of the new corpus.

Overall, the outcome of this data-curation process, in terms of the new corpus, led to significant modifications in three key areas:

1. **Format**. The new corpus is in a tabular format, rather than the EFCAMDAT's original XML format.
2. **Content**. The new corpus has been cleaned to remove texts containing issues that are likely to interfere with analyses relating to lexical development.
3. **Structure.** The new corpus is split into two samples, to account for some tasks containing groups of texts written in response to different prompts.

2.2 Preparing the new corpus

2.2.1 Selecting the sample

Because the new corpus was created with large-scale quantitative analyses of L2 lexical development in mind, the first step was to ensure that there were sufficient texts available for each combination of nationality and L2 proficiency level.[6] Accordingly, I selected only those nationalities and proficiency levels that had enough texts for my analyses. This is in line with many prior studies that used the EFCAMDAT. For example, Murakami (2016) and Shatz

[6] By "sufficient texts" I mean that the nationality generally had multiple texts available for each task at every L2 proficiency level, to ensure that there was a fairly full and balanced representation of all tasks, proficiency levels, and nationalities in the sample that I analyzed. At this stage, I decided to err on the side of including—rather than omitting—nationalities and proficiency levels that I was uncertain about, since it will be possible to select a sub-sample later if necessary, but not to analyze any data not included in this initial sample. Eventually, as discussed in Appendix B ("Sample information"), I omitted from my analyses the Turkish nationality and the C1 CEFR level (corresponding to EFCAMDAT levels 1–15), due to the low number of texts that they contained, which did not allow for full and balanced analyses.

(2019) examined only the top 10 nationalities with most texts in the EFCAMDAT, while Alexopoulou et al. (2015) and Geertzen et al. (2014) examined only the top 5.

In terms of proficiency level, texts from levels 1-15 were kept, while those from level 16 were omitted. There were relatively few texts at the omitted level (1,940, only 0.16% of the total), which were spread across multiple nationalities and tasks. In addition, levels 1-15 were grouped in bands of 3 based on EF's guidelines, while level 16 was on its own (Geertzen et al., 2013). Furthermore, level 16 was the only level listed as being above the maximum proficiency level set by several proficiency standards, such as the TOEFL.

In terms of nationality, texts from the 11 nationalities with most texts were kept: Brazilian, Chinese, Taiwanese, Russian, Saudi Arabian, Mexican, German, Italian, French, Japanese, and Turkish. These nationalities accounted for the vast majority of texts in the corpus (~89%), and there were relatively few texts spread out across the other 187 nationalities. Overall, 1,051,939 texts fit these criteria (89.1% of the texts in the EFCAMDAT).

2.2.2 Format: converting from XML to a tabular format

The EFCAMDAT was originally made available in XML format; a sample text with the original XML formatting appears in Figure 1.

```
<writing id="135" level="2" unit="5">
<learner id="73293" nationality="br"/>
<topic id="13">Making notes for a visitor</topic>
<date>2012-03-14 13:01:30.430</date>
<grade>96</grade>
<text>
Welcome to my house. Near the my house there is recreation center. Opposite to the
recreation center there is a soccer stadium. Between the recreation center and the soccer
stadium there is many restaurants. You guys enjoy!
</text>
</writing>
```

Figure 1. Sample text from the EFCAMDAT, with the original XML formatting

The EFCAMDAT was imported from XML format using R, together with the *XML* package and a custom function (Lang, 2020). This converted the texts and all their metadata into tabular *xlsx* format, where each row represents a single text. In addition, the following markup tags were modified, to clean the texts for analysis:

- 762,221 <*br/*> and 35 <*br*> tags were replaced with a space.
- 88,343 *&amp;quot;* tags were replaced with a single set of quotation marks.
- 6,872 *&amp;* tags were replaced with the word *and*.
- 1 <*/code*> tag was replaced with a space.

2.2.3 Content: analyzing and removing texts

2.2.3.1 Texts with problematic markup tags

A small number of texts contained the *&lt;* and *&gt;* markup tags, which stand for ‘<’ and ‘>’ respectively. Texts containing these tags were removed, because they were generally accompanied by problematic data, such as improperly formatted error tags provided by teachers, together with suggested corrections. This included, for example:

- &lt;&lt;&lt;&lt;IS&lt;correct&gt;/correct&gt;

- &lt;&lt;C, PU&lt;
- MY&lt;&lt;x&gt;y&lt;My).

These tags were not supposed to be in the version of the EFCAMDAT that was used here, which is meant to be free of annotations, and they would have interfered with future analyses, for example by inserting words into the text that the learner did not write. The reason why the full texts were removed is that the tags were inconsistent in terms of structure, so there was no simple scalable way to remove them while preserving the original texts they were in.

There were 4,554 *&lt;* tags and 1,631 *&gt;* tags in the sample, spread across only 1,329 texts (0.1% of texts at this stage). To remove them, two R packages were used: *stringr* to detect the relevant strings in texts (Wickham & RStudio, 2019), and *dplyr* (Wickham, François, et al., 2019) to filter texts based on the detected strings. After this removal, 1,050,610 texts remained.

2.2.3.2 Ultra-short texts

Ultra-short texts were defined as texts with fewer than 20 words, since such texts were below the minimal wordcount that learners were instructed to write, even at the lowest proficiency level. These texts often contained various issues. For example, many contained just random symbols (e.g. "???,??!??????,????????!" in text #876464) or only a few words (e.g. just "Hi," in text #613359). Similarly, there were over 20,000 such texts that were close variants of the same sentence ("Good evening. How are you. I'm fine, thanks. We're busy. Good night.").

Wordcounts were calculated using the *stringr* package (Wickham & RStudio, 2019) and a custom search pattern. 68,976 ultra-short texts were removed from the sample (6.6% of texts at this stage). Their average length was 13.49 words (*median* = 13, *standard deviation* = 3.38). Most of these texts (51,460, 74.6%) came from the first three tasks, with the majority (40,152) coming from the first one.

After this removal, 981,634 texts remained.

2.2.3.3 Non-English texts

Texts that were not written in English were removed. This included texts that contained gibberish of various forms, texts that were written entirely in a foreign language, and texts that contained substantial portions written in a foreign language. This problematic material often appeared due to technical issues, such as when the L1 instructions were copied into the text.

These texts were identified using the *cld2* library in R, which relies on a Bayesian classifier that identifies the language of texts (Ooms, 2018). The threshold for removal was the maximal proportion of English in the text (0.99), to ensure that the texts did not contain substantial portions of foreign-language material. Overall, only 16,925 texts containing significant levels of non-English text were removed (1.7% of texts at this stage). After this removal, 964,709 texts remained.

2.2.3.4 Duplicate texts

Duplicate texts were texts that were almost identical to each other in substantial portions. This generally occurred as a result of reusing source material from the task almost verbatim. For example, the texts in task #64 often had the exact same opening in response to the prompt "Claiming back your security deposit": "Dear Sir, I am writing to ask your advice about a problem I have with my landlord and the real estate agent…".

As with the other steps in the cleanup process, there are advantages and disadvantages to this removal. Specifically, the main advantage to removing these texts is that the direct reuse of source material could obscure L1 effects and other linguistic patterns in unpredictable ways. Conversely, the main disadvantage of removing these texts is that this could lead to the removal of some meaningful linguistic patterns, such as the use of formulaic language. However, this concern was mitigated, as this issue appeared to be relatively task dependent, rather than proficiency dependent. For example, the task with the highest proportion of duplicate texts (69.7%) was task #64, which is relatively advanced. This suggests that the issue of duplicate texts occurred, to a substantial degree, as a result of task effects and idiosyncrasies in the learning situation. In addition, the potential issues with this removal were further mitigated, as texts were removed only when they contained a substantial portion of identical, overlapping phrasing, down to letters and punctuation marks.

To calculate similarity between texts in the database, the *stringdist* package in R was used (Van der Loo, 2014). Specifically, the analysis used the *hamming* method, an edit-based algorithm that calculates the number of substitutions required to get from one string to another. To use it, trimmed versions of each text were created, which contained only the first 100 characters, since this method requires that the compared texts be of identical length. Texts were trimmed specifically to 100 characters, as the shortest text was 104 characters, and 100 represented a close and round number. This is beneficial when determining the similarity threshold later, and provides a proportion that is simple to replicate.

Then, to determine the threshold of similarity at which texts would be considered duplicates, an initial analysis was conducted on a sample of texts from Brazilian and Japanese learners in tasks #1 and #73. This sample was chosen as it represents two distinctly different nationalities and tasks, which contain different numbers of texts (16,229 and 3,513 for Brazilian, 737 and 307 for Japanese, in tasks #1 and #73 respectively).

A similarity matrix was calculated for the texts in this sample, and duplicate texts based on a similarity threshold of ‘5’ were extracted. This means that in cases where a trimmed text required fewer than five substitutions to be transformed into a different text in the sample, the two texts were designated as duplicates. Then, duplicate texts based on a similarity threshold of ‘10’ were also extracted, and the results between the two thresholds were compared manually by examining the list of new texts that were identified as duplicates, and checking whether they appeared to include true duplicates or false positives. This process was repeated, each time increasing the threshold by increments of 5 (10 → 15 → 20…). Eventually, 40 was identified as the optimal threshold, since it appeared to lead to the identification of new duplicates compared to a lower threshold of 35, and because increasing the threshold to 45 appeared to lead to a substantial increase in false positives.

Finally, a similarity matrix was calculated on the main sample, using ‘40’ as the threshold. Because each text must be compared against all other texts, this calculation involves potentially prohibitive computational complexity when run on large-scale datasets such as the EFCAMDAT. To resolve this, the analysis was run separately for each combination of nationality and task (for example, texts in task #1 among Japanese speakers, texts in task #1 among German speakers, etc.). This reduces the complexity of the calculation, and is unlikely to have a substantial impact on its outcome, since within-nationality duplicates are more likely than between-nationality duplicates, and since between-task duplicates are unlikely.

Based on this, 194,722 texts were removed (20.2% of the sample at this stage). Certain tasks were more likely to contain duplicate texts; for example, 8.9% of texts in task #23 were removed as duplicates, compared to only 5.1% of texts in task #53. Higher proficiency tasks were less likely to have texts marked as duplicates, but there were many cases where higher-level tasks had a higher proportion of duplicates than lower-level tasks (the correlation between proportion of duplicates per task and task number was *Spearman's rho*=−0.46, *p*<.001). After this removal, 769,987 texts remained.

2.2.3.5 Outlier texts based on wordcount

This step targeted texts that were anomalously short or long. Such texts often suffered from various issues, such as the inclusion of large amounts of irrelevant material, for reasons that are unclear. For example, text #455618 was anomalously long, with 129 words at task #1, where the average wordcount was 32, and contained a letter about a company's logo in response to the prompt "Introducing yourself by email".

Outlier texts in terms of wordcount were identified using *Tukey's method*. This means that, for each task, outlier texts were those that had a wordcount 1.5 *interquartile ranges* (IQR) below the 1st quartile or above the 3rd quartile of wordcounts for texts from the same task (Kannan et al., 2015). Accordingly, a different set of problematic texts were identified using this method compared to the one for removing ultra-short texts, since this method accounts for differences in wordcounts between tasks. For example, this means that text #1211137 was removed in this step, since it had a wordcount of 24 at task #26, where the average wordcount was 63. Note that it would have been insufficient to use only this method without first removing ultra-short texts, because of the low average wordcount in many of the low-proficiency tasks, especially when ultra-short texts are included.

Based on this analysis, 34,607 texts were removed (4.50% of the sample at this stage). Of these, 5,717 (16.52%) were short outlier texts and 28,890 (83.48%) were long outlier texts. After this removal, 735,380 texts remained.

2.2.4 Structure: classifying texts based on prompt

To explain this process, it helps to first define three terms:

- ***Task*:** this is the specific lesson that learners' texts are categorized under (e.g. "task #11"). Task numbers are listed sequentially in the EFCAMDAT, and range from 1-128, with 8 tasks per proficiency level.
- ***Prompt*:** this is the prompt that texts are written in response to (e.g. "Writing a weather guide for your city"). Each task has a corresponding prompt listed in the EFCAMDAT.

- ***Topic*:** this is the topic that a text revolves around (e.g. “weather”), as determined by classification software that will be described in this section.

Many texts in the EFCAMDAT did not correspond to their listed task prompt. The reason for this issue was as follows:

- Initially, each task was associated with a certain prompt. For example, task #11 had the prompt titled “Writing a weather guide for your city”.
- At some point, the prompts for some tasks were replaced with new ones. For example, the prompt for task #11 was changed to something such as “Describe people’s favorite sport in your country”.
- This change in prompt was *not* reflected in the database. Accordingly, all texts belonging to the same task number were listed together, regardless of which prompt they were written in response to. For example, task #11 contained texts written in response to the prompt on writing a weather guide, together with texts written in response to the new prompt on describing people’s favorite sport.

Accordingly, it was necessary to do the following:

- Determine which tasks contained groups of texts corresponding to multiple prompts.
- Determine how many prompts were used in such tasks.
- Categorize the texts in such tasks based on the prompt that they corresponded to.

Since no information regarding the different *prompts* was available in the database, it was necessary to find a scalable way to analyze the *topics* that texts revolved around. To do this, I first grouped texts from each task (e.g. task #1, task #2…), and used the *tm* package in R (Feinerer & Hornik, 2018) to create a document-term matrix, with the term frequencies for each text. Then, I used the *topicmodels* package (Grün & Hornik, 2011) to estimate a *latent Dirichlet allocation* (LDA) model using *Gibbs sampling*. For a visual representation of this process, see Figure 2.

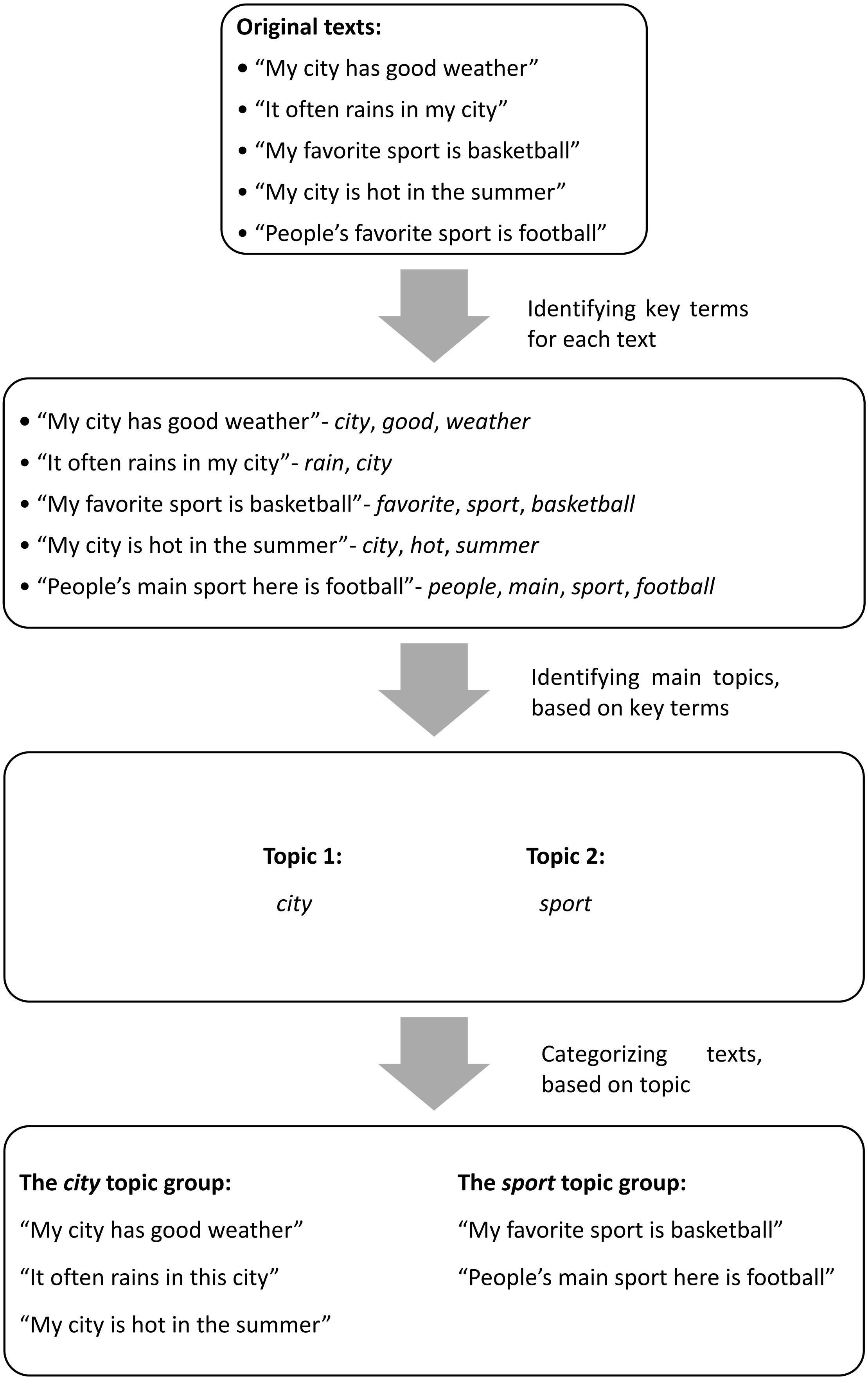

Original texts:
• "My city has good weather"
• "It often rains in my city"
• "My favorite sport is basketball"
• "My city is hot in the summer"
• "People's favorite sport is football"
Identifying key terms for each text
• "My city has good weather"- *city*, *good*, *weather*
• "It often rains in my city"- *rain*, *city*
• "My favorite sport is basketball"- *favorite*, *sport*, *basketball*
• "My city is hot in the summer"- *city*, *hot*, *summer*
• "People's main sport here is football"- *people*, *main*, *sport*, *football*
Identifying main topics, based on key terms
Topic 1:
city
Topic 2:
sport
Categorizing texts, based on topic
The *city* topic group:
"My city has good weather"
"It often rains in this city"
"My city is hot in the summer"
The *sport* topic group:
"My favorite sport is basketball"
"People's main sport here is football"

Figure 2. Rough illustration of the process used to classify texts based on topic

This process requires that the number of topics per task be specified in advance. Accordingly, to determine the appropriate number of topics, I started by running the process with two topics, and then tried increasing that number to three, while manually inspecting the texts. This revealed that the maximum number of prompts was ‘2’, as dividing texts into more than two topics led to groupings that were *not* based on a difference in prompt. For example, if texts written in response to the prompts “a weather guide for your city” and “people’s favorite sport in your country” were divided into more than two topics, then texts written in response to the same prompt would be separated; e.g. texts revolving around a weather guide might be split into those that primarily use keywords such as [*winter*/*cold*/*rain*] and those that use keywords such as [*summer*/*hot*/*sun*].

A single exception was task #13, where the classification software used the same keyword (‘there’) to classify texts from both topics. Accordingly, I re-ran the analysis for this task with three topics in the LDA model. I then examined the texts and combined two of the topics (under the keywords ‘there’ and ‘house’), while the third topic (under ‘neighborhood’) was marked as corresponding to a different prompt.

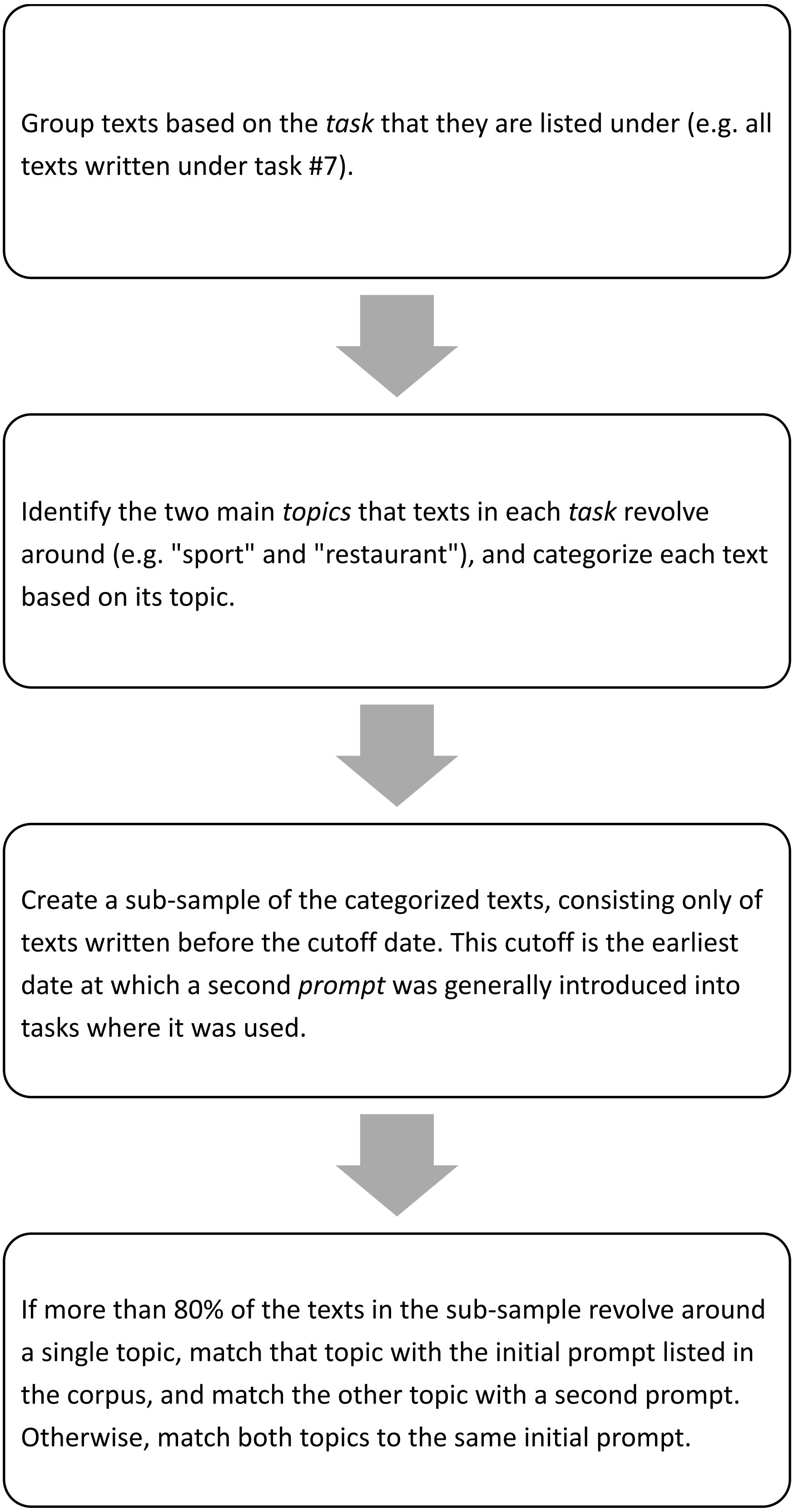
Next, it was necessary to determine which tasks contained groups of texts corresponding to two prompts, and then classify texts accordingly. An examination showed that, in tasks with texts corresponding to two prompts, texts were initially written in response to the first prompt, until a certain date when the new prompt replaced the first. Accordingly, a sub-sample of the corpus was created, containing only texts submitted before 2012-07-04, which was established as the earliest approximate point when the second prompt was introduced.

Then, the topics of the texts in the sub-sample were analyzed separately for each task:

- In cases where most texts (80%+) before the cutoff date belonged to a single topic, the task was categorized as having two prompts. Essentially, if most texts before the cutoff revolved around a single topic, this indicated that the topic corresponded to an initial prompt, while the less frequent topic corresponded to a second prompt that was introduced only after the cutoff. For example, if almost all of the texts before the cutoff revolved around the topic *city*, and almost none revolved around the topic *sport*, then it was likely that texts written about *sport* were based on a second prompt, which was introduced later.

- In cases where fewer than 80% of texts before the cutoff belonged to a single topic, the task was treated as having a single prompt. Essentially, if the texts before the cutoff date revolved around two topics in relatively similar proportions, then there was likely only one prompt for the task, since the similarity in proportion indicated that the division into topics was *not* based on a difference in prompt. For example, if texts before the cutoff revolved around the topics *restaurant* and *food* in relatively similar proportions, then it was likely that the texts were written in response to the same prompt, and that they simply used slightly different keywords.

One concern was that there might be tasks where one topic was much more common overall. However, this was ruled out, given that the most extreme ratio between topics in the full sample was 2.3:1 (between the second and the first topics in task #92), and the overall mean ratio between the first and second topics was 0.89 (*median* = 0.86, *SD* = 0.29). Conversely, the cutoff point used to determine whether two prompts were used was the much higher ratio of 4:1 (i.e., 80%). Overall, the procedure used to classify texts is outlined in Figure 3.



Group texts based on the *task* that they are listed under (e.g. all texts written under task #7).

Identify the two main *topics* that texts in each *task* revolve around (e.g. "sport" and "restaurant"), and categorize each text based on its topic.

Create a sub-sample of the categorized texts, consisting only of texts written before the cutoff date. This cutoff is the earliest date at which a second *prompt* was generally introduced into tasks where it was used.

If more than 80% of the texts in the sub-sample revolve around a single topic, match that topic with the initial prompt listed in the corpus, and match the other topic with a second prompt. Otherwise, match both topics to the same initial prompt.

Figure 3. Outline of the process that was used to identify and classify texts written in response to different prompts

In cases where all of the texts from a given task were established as having been written under the same prompt, they were all kept in the sample (31 tasks, 25.8% of total). Conversely, in cases where texts from a given task were established as having been written under two prompts, only texts written using the initial prompt were kept in the main sample (89 tasks, 74.2%). Accordingly, 329,318 texts (44.78% of texts) were designated as having been written in response to a second prompt, and were consequently separated into a second sample.

Finally, the texts in the second sample were further cleaned. This involved removing texts that were categorized as having been written in response to the second prompt despite being written before the cutoff date, which was the earliest point when the new prompt was generally introduced. The cutoff date used at this stage was 2013-04-03, which was later than the cutoff used previously. This is because the second prompt was often introduced around this later date, so using it allowed for the removal of more irrelevant texts. This led to the removal of 12,098 texts (3.67%), leaving 406,062 texts in the first sample and 317,220 texts in the second sample.

An important limitation of the second sample is that it does not list the prompts that learners responded to in their texts, since such data is not available in the EFCAMDAT. However, the original prompts from the first sample are still listed in the second sample, to maintain continuity between the samples; this ensures that the two samples share the same data structure, which means that researchers can easily concatenate them into a single sample if they wish. Nevertheless, it is possible to estimate the prompts manually, by reading the texts. Alternatively, it is possible to identify the key topics that the texts revolve around, by using the same keyword-extraction method that was implemented earlier; one such keyword is already listed for each text in the new version of the corpus, based on the earlier extraction process.

2.3 Discussion and conclusion

Overall, this report outlined a comprehensive process used to modify and refine a large-scale English learner database—the EFCAMDAT—in terms of its format, content, and structure. The process used to create the derivative corpus is outlined in Figure 4.

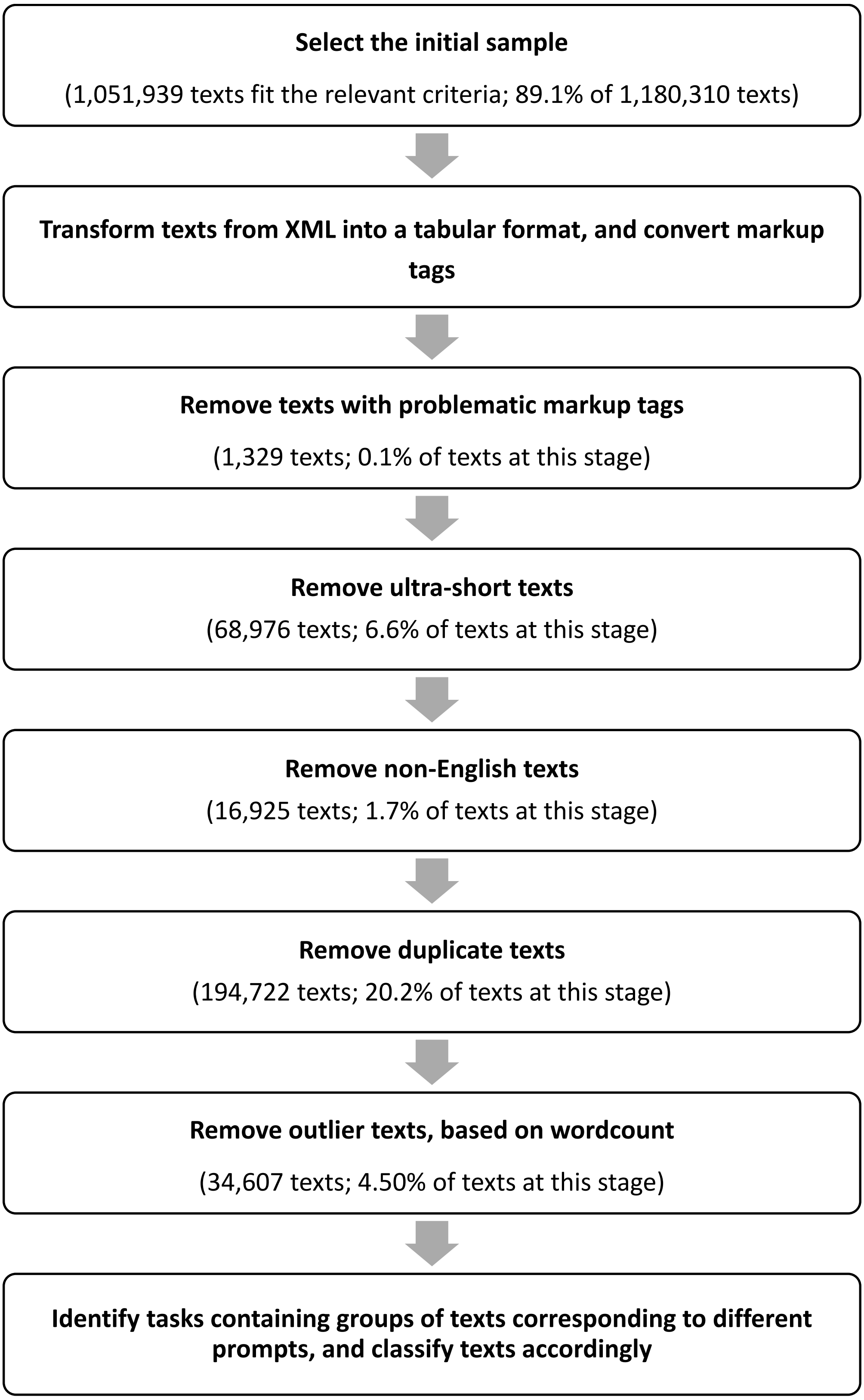

Select the initial sample
(1,051,939 texts fit the relevant criteria; 89.1% of 1,180,310 texts)
Transform texts from XML into a tabular format, and convert markup tags
Remove texts with problematic markup tags
(1,329 texts; 0.1% of texts at this stage)
Remove ultra-short texts
(68,976 texts; 6.6% of texts at this stage)
Remove non-English texts
(16,925 texts; 1.7% of texts at this stage)
Remove duplicate texts
(194,722 texts; 20.2% of texts at this stage)
Remove outlier texts, based on wordcount
(34,607 texts; 4.50% of texts at this stage)
Identify tasks containing groups of texts corresponding to different prompts, and classify texts accordingly

Figure 4. Summary of the preparation process of the corpus

Based on this, from an initial database containing 1,180,310 texts, a corpus was created with 406,062 texts (~24,826,000 word tokens) in the first sample and 317,220 texts (~20,564,000 word tokens) in the second sample. These samples cover 120 and 89 topics respectively, and contain texts written by learners from 11 nationalities and with a large range of English proficiency levels (CEFR A1-C1). The numbers of texts per nationality and CEFR level in the new samples are listed in Table 1.

Table 1. Number of texts in the derivative corpus, per nationality and CEFR level. Nationalities are listed by total number of texts in the first sample, in decreasing order

Nationality	Number of texts (first sample)						Number of texts (second sample)					
	Total	A1	A2	B1	B2	C1	Total	A1	A2	B1	B2	C1
Brazilian	149,297	75,497	45,407	20,989	5,830	1,574	164,241	85,191	42,105	25,520	9,412	2,013
Chinese	86,660	45,008	29,318	10,321	1,763	250	20,317	10,494	6,021	2,730	936	136
Mexican	34,559	19,296	9,847	4,102	1,114	200	30,204	15,998	7,645	4,500	1,740	321
Russian	32,243	12,295	10,885	6,329	2,066	668	17,078	7,249	4,652	3,449	1,443	285
German	24,705	8,041	7,860	5,051	2,698	1,055	16,717	4,652	4,487	4,083	2,669	826
French	19,135	7,626	6,253	3,688	1,242	326	13,384	4,610	3,755	3,188	1,528	303
Italian	18,959	5,899	6,832	4,291	1,466	471	16,469	5,046	5,010	4,166	1,749	498
Saudi Arabian	13,152	7,463	3,729	1,412	417	131	16,156	8,089	4,874	2,301	727	165
Taiwanese	11,711	4,116	4,298	2,506	650	141	10,900	3,668	3,731	2,490	893	118
Japanese	9,149	3,337	3,095	1,903	640	174	7,937	2,812	2,409	1,837	701	178
Turkish	6,492	3,085	2,067	914	301	125	3,817	1,683	1,064	769	253	48
Total	*406,062*	*191,663*	*129,591*	*61,506*	*18,187*	*5,115*	*317,220*	*149,492*	*85,753*	*55,033*	*22,051*	*4,891*

As noted earlier, the new corpus was created to facilitate large-scale quantitative analyses of L2 lexical development, using the data available in the EFCAMDAT. Accordingly, some of the procedures in the data-curation process may not be appropriate for other types of analyses; a notable example of this is the removal of duplicate texts, which could be an issue for analyses that focus on formulaic language. As such, to facilitate the use of the EFCAMDAT for other purposes, in addition to making the final version of the new corpus available, I have also made available additional versions of the corpus from different steps of the data-curation process, together with the key R scripts that I used. All these materials, together with other relevant ones, such as a glossary of variables, are available on the official EFCAMDAT site (https://corpus.mml.cam.ac.uk/). They are currently listed there under the "Resources" page, as the "EFCAMDAT Cleaned Subcorpus".

In addition to introducing the new derivative corpus, this report can also inform future work on other learner corpora, by identifying issues that researchers may encounter during data curation and analysis, and by proposing scalable solutions that they may use. This is something that is becoming increasingly necessary, given the growing use of large-scale learner corpora that are based on educational and social platforms, and that were therefore not originally collected with research in mind.

3 Study 1: The influence of crosslinguistic similarity on L2 lexical diversity

This chapter outlines a study that examines how crosslinguistic lexical similarity at the language level influences the lexical diversity of L2 texts. In doing so, it also provides valuable insights into lexical diversity, in terms of factors such as how it develops as learners' L2 proficiency increases, how much it varies between and within L2 proficiency levels, and how strongly it is influenced by task effects.

The abstract of the paper is as follows:

> We examined the potential influence of L1-L2 lexical similarity on L2 lexical diversity using two matching sub-corpora, containing 8,500 and 6,390 English texts, written by speakers of 9 typologically diverse L1s, in the A1–B2 CEFR range of L2 proficiency. Lexical similarity did not influence L2 lexical diversity, at any proficiency level. This suggests that the facilitative effect of lexical similarity that is found in processing and broad learning outcomes does not necessarily extend to L2 production, at least in the case of certain global measures, such as lexical diversity, and certain task-based settings, where lexical choices are driven primarily by the constrained communicative needs of the tasks. This is supported by the strong task effects that were found, which emerge as an important factor influencing vocabulary use that needs to be taken into account when assessing lexical diversity. Additionally, lexical diversity was shown to be a useful indicator of L2 proficiency, but primarily on group data, given the substantial variability involved, and from beginner to intermediate levels, when it starts to plateau.

Note that, since this chapter is written in the format of a paper, you will have already encountered some of the material here previously in the thesis, especially when it comes to the research background and sample information.

This work was done in collaboration with Dora Alexopoulou and Akira Murakami.

3.1 Introduction

[Due to space constraints—and to avoid repetition—I removed the opening of this introduction, which largely repeats the information from the thesis's introduction (§1.1 and §1.2) on transfer, crosslinguistic influence, crosslinguistic similarity, and cognate facilitation. Instead, we turn immediately to the discussion of the key study by Crossley and McNamara, in the context of which we introduce lexical diversity.]

…Indeed, Crossley and McNamara (2011) suggest that [the facilitative effect of crosslinguistic similarity] may not extend to task-based settings, at least when it comes to certain aspects of L2 production. Specifically, their research examined a sample of 599 L2 English texts from the *International Corpus of Learner English*, written in response to one of few prompts for argumentative-style essays, by speakers with high-intermediate to advanced L2 English proficiency, and with Czech, Finnish, German, or Spanish as an L1. They found that learners' L1 had no effect on several global L2 lexical measures, including *lexical diversity*, which is the range of different words that are used in a text, where a higher range indicates greater diversity (McCarthy & Jarvis, 2010).

Lexical diversity captures the rate of word repetition in texts, and learners with a larger vocabulary tend to repeat words less on average. As such, this measure is indicative of learners' L2 vocabulary and of their ability to use it, and it generally increases as learners' L2 proficiency increases, so it is often used in language assessment, though the correlation between lexical diversity and L2 proficiency is imperfect (Alexopoulou et al., 2017; Crossley et al., 2015; Hout & Vermeer, 2010; Jarvis, 2013; Johnson, 2017; Kyle et al., 2021; McCarthy & Jarvis, 2010; Murakami & Alexopoulou, 2016; Treffers-Daller et al., 2018; Yan et al., 2020; Zenker & Kyle, 2021). Lexical diversity is also significantly influenced by task effects, so writing a résumé, for example, could elicit different average lexical diversity than describing the plot of a movie, which supports the suggestion that task effects may override the influence of lexical similarity on lexical diversity (Alexopoulou et al., 2017; Johnson, 2017; Torruella & Capsada, 2013; Yu, 2010; Zenker & Kyle, 2021).

However, Crossley and McNamara (2011) mention that their research is limited in terms of the scope of their sample and their analyses. This casts some uncertainty regarding the lack of L1 effect on L2 lexical diversity that they found, particularly in light of the extensive findings on the associated L1 effects when it comes to L2 processing, acquisition, and word choice.

Here, we will address the key limitations that they mention, and follow their call to replicate and extend their research. We will do so by analyzing the influence of learners' L1 on their L2 lexical diversity using a substantially broader and more diverse sample, in terms of the number of learners, number of texts, number and type of tasks, number and diversity of L1s, and range of L2 proficiency levels. In addition, our models will control for more variables, including L1-L2 lexical similarity, which they did not consider in their analysis, as well as L2 proficiency and task. The use of such sample and models, together with our focus on lexical diversity in our analyses, will also help shed light on the influence of *L2 proficiency* and *task* on lexical diversity, which is necessary given that past studies that examined these effects generally did not focus on them and/or used narrower samples, and is important given the common use of lexical diversity in language assessment.

Based on this, our research questions are as follows:

1. Does L1-L2 lexical similarity influence L2 lexical diversity in task-based settings, in general or at early proficiency levels? This is the main research question, and the findings of Crossley and McNamara (2011) suggest that it might not, despite the robust L1 effects found for similar phenomena, such as L2 processing and acquisition.
2. How does L2 proficiency influence lexical diversity? We expect lexical diversity to increase with L2 proficiency, but this correlation is likely to be imperfect, particularly as the increase in lexical diversity might slow down as learners' L2 proficiency increases (Alexopoulou et al., 2017; Treffers-Daller et al., 2018).
3. To what extent does *task* influence lexical diversity? We expect task effects but are uncertain regarding their magnitude, so investigating this in a broad sample will help us understand how these effects relate in magnitude to the effect of L2 proficiency.

3.2 Methodology

3.2.1 Learner sample

Here, we present the key details about our learner sample. For more information about it, see the supplementary "Sample information" document [included in the thesis as Appendix B].

The learner sample came from the *EF-Cambridge Open Language Database* (EFCAMDAT), an open-access L2-English learner corpus, containing texts written by learners in EF's online English school (Geertzen et al., 2013; Y. Huang et al., 2018). When a learner joins EF's school, their starting proficiency level is determined using a dedicated placement test. The EFCAMDAT spans 16 proficiency levels that EF aligned with common proficiency standards, such as the *Common European Framework of Reference for Languages* (CEFR). Each level consists of several lessons; after completing a lesson, learners are assigned a writing task that they submit online, and that they then receive feedback on from a teacher. These tasks (discussed in more detail in the supplementary "Sample information"), cover a wide range of styles and topics, such as describing your favorite day, reviewing a song, writing an online profile, or giving instructions to a house-sitter. The curriculum is standardized, so learners with different L1s follow the same lessons and activities, and are given the same writing tasks.

We used the *EFCAMDAT Cleaned Subcorpus*, which is available on the EFCAMDAT site (https://corpus.mml.cam.ac.uk/) and which is outlined in detail in Shatz (2020). The key feature of this dataset is that it is split into two sub-corpora, each containing texts written by similar learners in response to different prompts. This means, for example, that both the first and the second sub-corpora contain texts written by German learners in task #4, but the learners in the first sub-corpus wrote their texts in response to a different prompt than learners in the second sub-corpus. As such, using this dataset presents two important advantages for research. First, it allows us to accurately categorize texts based on the task that they belong to. Second, as noted by Shatz (2020), this offers an opportunity to conduct our analyses on two similar but distinct learner samples, which serves as a form of replication.

Texts were selected from this dataset, in a random manner but kept relatively balanced across L1s, proficiency levels, and tasks, as outlined in the supplementary "Sample information" document. The final samples are listed in Table 2.

Table 2. Final learner samples.

L1s (nationalities)	Arabic (Saudi Arabian), French, German, Italian, Japanese, Mandarin (Chinese), Portuguese (Brazilian), Russian, Spanish (Mexican)
Target L2	English
L2 proficiency levels	EFCAMDAT 1–12 (equivalent to CEFR A1–B2) [a]
Tasks per corpus	95 (first) / 71 (second) [b]
Number of texts per L1 per task	10 [c]
Total number of texts (corpus)	8,500 (first) / 6,390 (second) [d]

[a] Every 3 EFCAMDAT levels correspond to a single CEFR level.
[b] There are 8 tasks per EFCAMDAT level in the first corpus, and 6 tasks per level in the second. One exception is task #51, in which texts from both corpora were classified under the first corpus due to limitations in the classification scheme, so we removed this task was removed from both corpora.
[c] In the first corpus, there are a few exceptions to this (14 out of 855, 1.64%), which had 2–9 texts (mean = 6.43, standard deviation = 1.79), as listed in the "Sample information" document.
[d] A few (~5%) of these texts were later removed as outliers (see §3.2.4).

3.2.2 Lexical distance

We calculated lexical distance[7] between the L1s and English using *Swadesh lists* from the *Automated Similarity Judgment Program* (ASJP), which are wordlists containing concepts that appear in nearly all languages, such as *water*, *full*, and *hear* (Swadesh, 1950; Wichmann et al., 2018). This source is often used to calculate lexical distance between languages, and has been validated extensively, as discussed in the supplementary information, for example through the comparison of distances based on it to distances based on expert judgments (Bakker et al., 2009; Schepens, van der Slik, et al., 2013b; Wichmann et al., 2010).

The Swadesh lists in the ASJP focus on a subset of 40 concepts, representing the most stable elements from the original list for language classification (Bakker et al., 2009; Holman et al., 2008b). To control for variations in completeness of Swadesh lists across languages, only the 38 concepts that are shared across all the L1s in the present sample were included. In addition, the ASJP's phonetic script was converted to IPA using the *asjp* library in Python (Sofroniev, 2018).

We calculated crosslinguistic lexical distance using *Levenshtein distance normalized* (LDN). This measure, which can be calculated in an automated, objective, and replicable manner for a large number of words from different languages, is a common measure of lexical distance, and has been extensively used and validated in previous studies, as shown in detail in

[7] See the supplementary information for an explanation why we use the term *lexical distance*.

the supplementary information [included in the thesis as Appendix D]. This includes research showing that it strongly correlates with expert cognancy judgments (Schepens, van der Slik, et al., 2013b), with distances based on morphological features (Schepens et al., 2020), and with various psycholinguistic measures, such as perceived language distance (Heeringa & Prokić, 2018). This also includes L2 acquisition research that used this measure to quantify crosslinguistic similarity, and found that it predicts L2 outcomes such as word knowledge (De Wilde et al., 2020) and overall L2 proficiency (Schepens et al., 2020; Schepens, van der Slik, et al., 2013b).[8]

LDN is calculated by taking the minimum number of character substitutions, additions, and deletions that are needed to transform one string to another, and then dividing this number by the length of the longer string, to account for variations in word length. For example, in the case of the word *knee*, the English-German pair /ni/ and /kni/ has an LDN of 0.33, since there is 1 character transformation (a /k/ is inserted or deleted), and the length of the longer string is 3. Here, we first calculated lexical distance between each L1 entry and its corresponding L2 (English) entry. Then, we calculated the overall L1-L2 lexical distance between each L1 and English, based on the mean distance of all the L1-L2 word pairs in that L1.

The results of the lexical distance calculations are presented in Table 3. The wide range of distances from English highlight the diversity of the L1s in our sample. The distances are largely in line with what is expected based on general language classification, with the Germanic and Romance L1s being the closest to English.

Note that arguments could be made for other types of distances, but as noted above, and as shown in detail in the supplementary information, this particular distance has been extensively validated, and is strongly correlated with other distances (both lexical and otherwise), so it is unlikely that the use of a different measure would substantially influence

[8] However, it is also important to note the limitations of using Levenshtein distance for language classification. One such key limitation is that it incurs the same cost for all character transformations, which does not always reflect underlying linguistic patterns accurately. For example, the English word “fish” /fɪʃ/ has an LDN of 1 from both its Spanish translation (“pez” /pes/) and its Hebrew translation (“דג” /dag/), despite being closer phonologically and etymologically to the Spanish translation, with which it could also be considered cognate. Another such limitation is that LDN looks only at formal similarity across words, but other factors (e.g., semantic and morphological similarity) may also play a role in language similarity, including at both the word level and the language level; this and other criticisms are also discussed in more detail in Greenhill (2011). In addition, the studies that validated the use of this measure used various different methodologies for different purposes, which do not always align with those of the present study, and these studies also likely had various limitations and shortcomings. As such, it is important to be cautious in interpreting language classifications that are based on this measure, and it will be beneficial to replicate results that use them with other measures of distance, as we do ourselves with a binary distance measure.

the results. This is further supported by the particular findings of our study (especially the estimated marginal means in Table 5), which suggest that variation in the exact distances that were calculated would not substantially change our key findings. Moreover, this is supported, as shown later, by supplementary models that we built, with a binary measure of lexical distance (based on whether the L1 is Indo-European like English), which fully replicated our findings with lexical distance based on the Swadesh lists.

Table 3. Lexical distance between each L1 and English, based on the phonological normalized Levenshtein distance from the closest synonym in the Swadesh lists. L1s are ranked in increasing order of mean distance.

	Lexical distance				
L1	mean	SD	median	IQR	range
German	.665	.27	0.67	0.50-0.92	0.00-1.00
Italian	.820	.20	0.83	0.72-1.00	0.29-1.00
French	.855	.19	1.00	0.75-1.00	0.25-1.00
Spanish	.862	.19	1.00	0.75-1.00	0.29-1.00
Portuguese	.878	.18	1.00	0.80-1.00	0.50-1.00
Russian	.883	.18	1.00	0.80-1.00	0.00-1.00
Japanese	.907	.14	1.00	0.83-1.00	0.50-1.00
Arabic	.916	.12	1.00	0.80-1.00	0.50-1.00
Mandarin	.922	.12	1.00	0.85-1.00	0.50-1.00

Note. Each L1 contained words corresponding to 38 unique meanings in English. The number of entries varied slightly between L1s, due to different numbers of L1 synonyms (the mean number of entries per L1 was 41, median = 39, SD = 3.67, range = 38–49).

3.2.3 Lexical diversity

Lexical diversity can be based on various measures, the simplest and best-known of which is the type-token ratio (TTR), which represents the number of *types* (unique words in the text), divided by the number of *tokens* (all the words in the text, regardless of repetition) (Torruella & Capsada, 2013). However, because the basic TTR measure is highly sensitive to text length, other measures have been developed from it (Covington & McFall, 2010; Fergadiotis et al., 2015; Granger & Wynne, 1999; Hout & Vermeer, 2010; Jarvis, 2013; Kyle et al., 2021; McCarthy & Jarvis, 2010; Michel, 2017; Torruella & Capsada, 2013; Zenker & Kyle, 2021).

Here, we assessed lexical diversity using the *measure of textual lexical diversity* (MTLD), which is described as follows:

> MTLD is an index of a text's LD [lexical diversity], evaluated sequentially. It is calculated as the mean length of sequential word strings in a text that maintain a given TTR value (here, .720). During the calculation process, each word of the text is evaluated sequentially for its TTR. For example, . . . of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people (.667) for (.714) the (.625) people (.556) . . . and so forth. However, when the default TTR factor size value (here, .720) is reached, the factor count increases by a value of 1, and the TTR evaluations are reset. Thus, given the previous example, MTLD would execute . . . of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people (.667) |||FACTORS = FACTORS + 1||| for (1.00) the (1.00) people (1.00) . . . and so forth.
>
> A partial factor value is calculated for the lexical remainders of a text (i.e., the final words that do not form a full factor). For example, a TTR of .887 forms 40.4% of the range between 1.00 and the full factor of .720. If a text contains 4 full factors and a remainder that has a TTR of .887, then the final factor count is 4.00 + 0.404 = 4.404…
>
> The total number of words in the text is divided by the total factor count. For example, if the text = 340 words and the factor count = 4.404, then the MTLD value is 77.203. Two such MTLD values are calculated, one for forward processing and one for reverse processing. The mean of the two values is the final MTLD value.
>
> (McCarthy & Jarvis, 2010, pp. 384–385)

We chose MTLD for several reasons. First, there is substantial prior research on it, which facilitates the interpretation of our findings and their comparison with those of others (especially Treffers-Daller et al., 2018). Furthermore, prior research shows that MTLD strongly correlates with other common measures of lexical diversity, such as *vocd-D*, *HD-D*, and *Maas* (Fergadiotis et al., 2015; McCarthy & Jarvis, 2010; Treffers-Daller et al., 2018), so findings that are based on it are reasonably generalizable.

In addition, MTLD is relatively robust to short texts and to variations in text length, compared to most other measures of lexical diversity (Fergadiotis et al., 2015; Koizumi, 2012; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020; Yan et al., 2020; Zenker & Kyle, 2021). However, because the shorter the text is the greater role the remainder plays in the calculation of MTLD, and because remainders represent approximations of lexical diversity, MTLD is less robust on short texts below a certain point (McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020). The lower bound for using MTLD has traditionally been 100 words, but comprehensive recent research by Zenker and Kyle (2021) examined the use of MTLD in texts as short as 50 words, and found that it is fairly robust there too.

Though some of the texts that we analyzed were below this lower bound, this potential issue did not appear to invalidate our findings, as shown in the results sections, and as expanded upon in the supplementary information. This is based on several pieces of evidence, including, most notably, that our key findings replicate when we analyze appropriate sub-samples with texts that are longer than 50 or 100 words. Furthermore, our results replicate when we conduct our analyses using another robust measure of lexical diversity—*moving-average type–token ratio* (MATTR)—which is calculated in a different way than MTLD (without remainders), and may therefore be more robust in short texts (Covington & McFall, 2010; Fergadiotis et al., 2015; Vidal & Jarvis, 2020; Zenker & Kyle, 2021). In addition, prior research supports the reliability of MTLD in the EFCAMDAT, including in shorter texts, by showing that it strongly correlates with measures of syntactic complexity (Alexopoulou et al., 2017).

Note that we calculated MTLD using the spelling-corrected texts, since spelling errors can inflate it; for more details on this and our technical approach, see the supplementary information.

3.2.4 Data analysis

We built mixed-effects linear regression models (Hox et al., 2018; Winter, 2019) for each corpus in our sample, with the following structure:

1. **Response variable**: *lexical diversity*, based on the MTLD of each text. In addition, we also built supporting models with MATTR as the response variables (see the supplementary information).
2. **Predictors**:
    a. *Lexical distance*, based on LDN. In addition, we built supporting models with a binary measure of lexical distance, based on language family, as explained in the supplementary information.
    b. *L2 proficiency*, based on EFCAMDAT proficiency level (1–12, corresponding to CEFR A1–B2) of the learner at the time they wrote the text (each lesson/task is classified under a certain proficiency level).

   c. *Interaction term* between the predictors, to determine whether the effect of *lexical distance* varies as a function of *L2 proficiency* (i.e., whether the effect of lexical distance weakens as L2 proficiency increases), since prior research suggests that the expected L1 effects are generally stronger at lower proficiency levels.

3. **Random effects** (random intercepts)[9]:
   a. *Learner*, to control for cases where learners had more than one text in the sample.[10]
   b. *Task*, as a categorical variable based on a unique identifier for each task. This allows us to control for *task effects*, which we operationalize here as the combination of all aspects of each writing task that might influence lexical diversity, aside from its associated L2 proficiency level (which we control for using the relevant predictor); this includes aspects such as the task's style and prompt, though our approach does not allow us to determine which aspects of the task are responsible for the task effects.[11] Note that the use of mixed-effects models allows us to assess such task effects despite the fact that each task is associated with only a single proficiency level (Hox et al., 2018; Winter, 2019), and this type of mixed-effects structure—where each group in a random grouping variable always takes the same potentially unique value along a continuous predictor—is conventional in both corpus linguistics (e.g., Levshina, 2018) and psycholinguistics (e.g., Vandenberghe et al., 2021).
   c. *L1*, as a categorical variable that controls for effects from the learners' L1 and their associated (e.g., cultural) background, beyond the effects of lexical distance. This use of *L1* as a random effect is similar to the use of *task* as a random effect, as outlined above.

Before building the models, we centered the predictors to reduce potential collinearity with the interaction term. In addition, we identified and removed ~5% of texts that were classified as outliers based on their MTLD, as discussed in the supplementary information, leaving 8,081 texts in the first corpus and 6,129 in the second. We also checked the statistical assumptions of

[9] Several models with random slopes were considered, but did not converge or were less optimal than the random-intercepts model, though their inclusion did not substantially influence our findings; more details on this appear in the supplementary information.

[10] The mean number of texts per learner after the removal of outliers was 1.36 in the first corpus and 1.41 in the second. For more information, see the "Sample information" document.

[11] Our operationalization of tasks is therefore distinct from most notions of task within task-based learning and teaching approaches (Alexopoulou et al., 2017), and we do not make a claim regarding the impact of any specific aspect of tasks.

the models and found no substantial issues, as shown in the supplementary information, and compared these models with baseline models with no lexical distance.

All the data and code that were used in the study are available in the following *Open Science Framework* (OSF) repository: https://doi.org/10.17605/OSF.IO/95HWB

3.3 Results

Figure 5 and Table 4 show that learners' lexical diversity (MTLD) generally increases as their L2 proficiency increases, though this increase appears less steep going from the B1 to the B2 CEFR level, particularly in the second corpus. In addition, this figure and table show there is substantial variability in MTLD both within tasks and within proficiency levels, based on the large standard deviation (SD) within each task/level, particularly compared to the differences between them. Together, this shows that lexical diversity is correlated with L2 proficiency, but this correlation is imperfect and involves substantial variability.

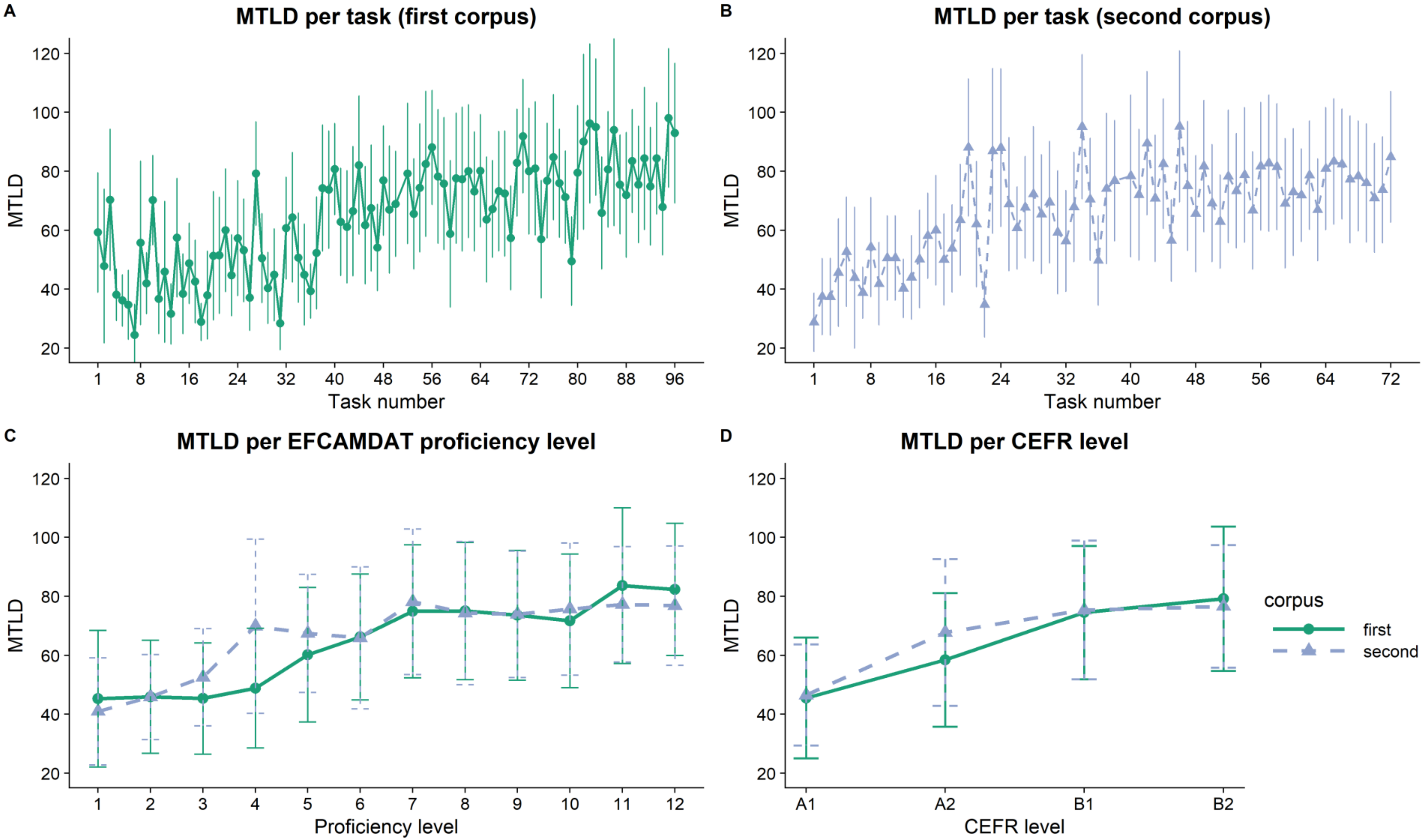


Figure 5. Mean lexical diversity (MTLD); error bars indicate one standard deviation. Listed per *task* in (A) and (B), per *EFCAMDAT proficiency level* in (C), and per *CEFR level* in (D). There are 8 tasks per EFCAMDAT proficiency level in the first corpus and 6 tasks per EFCAMDAT level in the second. There are 3 EFCAMDAT levels per CEFR level in both corpora. Raw correlations appear in the supplementary information.

Table 4. Mean lexical diversity (MTLD) per CEFR (L2 proficiency) level, corresponding to Figure 5D.

	First corpus			Second corpus		
CEFR	n	mean MTLD (*SD*)	MTLD increase (percentage)	n	mean MTLD (*SD*)	MTLD increase (percentage)
A1	1921	45.53 (20.49)	-	1508	46.56 (17.13)	-
A2	2088	58.49 (22.68)	12.96 (28.46%)	1553	67.76 (24.86)	21.20 (45.54%)
B1	2042	74.52 (22.63)	16.03 (27.41%)	1500	75.36 (23.52)	7.60 (11.22%)
B2	2030	79.20 (24.50)	4.68 (6.29%)	1568	76.62 (20.81)	1.25 (1.66%)

Note. *n* denotes the number of texts. *MTLD increase* is the mean MTLD at that CEFR level, minus the mean MTLD of the previous level. The *percentage* increase is the mean MTLD at that CEFR level, divided by the mean MTLD of the previous level, minus 1 and then times 100. Calculations are based on unrounded values.

Figure 6 shows there is much similarity between the L1s in terms of mean MTLD and in terms of MTLD increase over L2 proficiency. The scatterplots and linear models in Figure 7 show that lexical distance does *not* predict MTLD at any CEFR proficiency level (as they all had a non-significant R^2 that is lower than .001). Furthermore, Table 5 shows that the different L1s had similar mean MTLD, and more importantly, that there is no association between lexical distance and MTLD, as the ranking of L1s in terms of lexical distance does *not* influence their ranking in terms of MTLD.

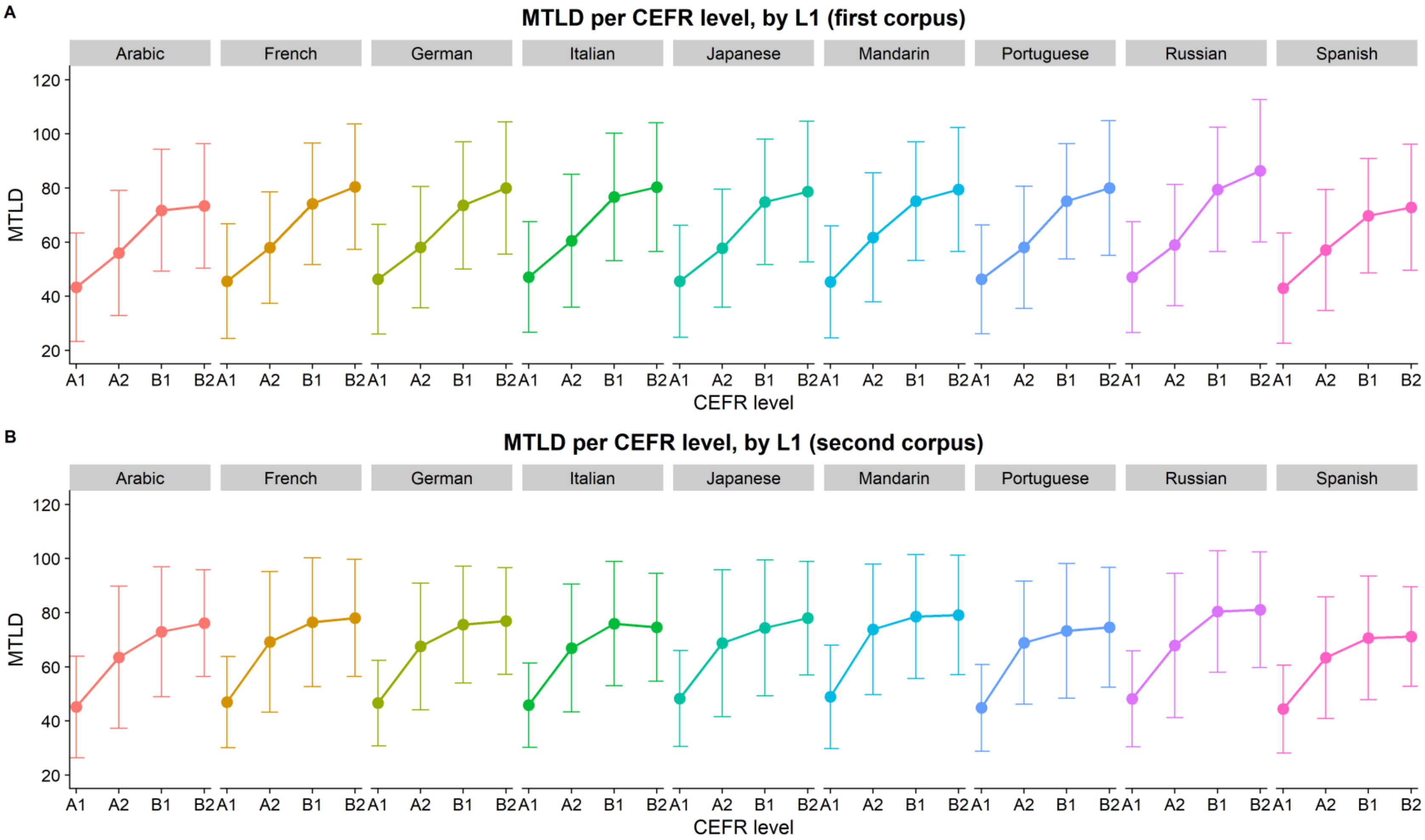


Figure 6. Mean lexical diversity (MTLD) per CEFR (L2 proficiency) level, by L1, in each corpus. Error bars denote one standard deviation. The exact values for these plots appear in the supplementary information. Note that the increase in MTLD is more linear in the first corpus, where there is a similar increase for A1→A2 and A2→B1, than in the second corpus, where there is a much steeper increase for A1→A2 than A2→B1. However, the initial values (at A1) and final values (at B2) are similar in both corpora, and in both there is a relative plateau for B1→B2, though this is more pronounced in the second corpus (where there is even a decrease for Italian, likely due to the plateau combined with noise in the data).

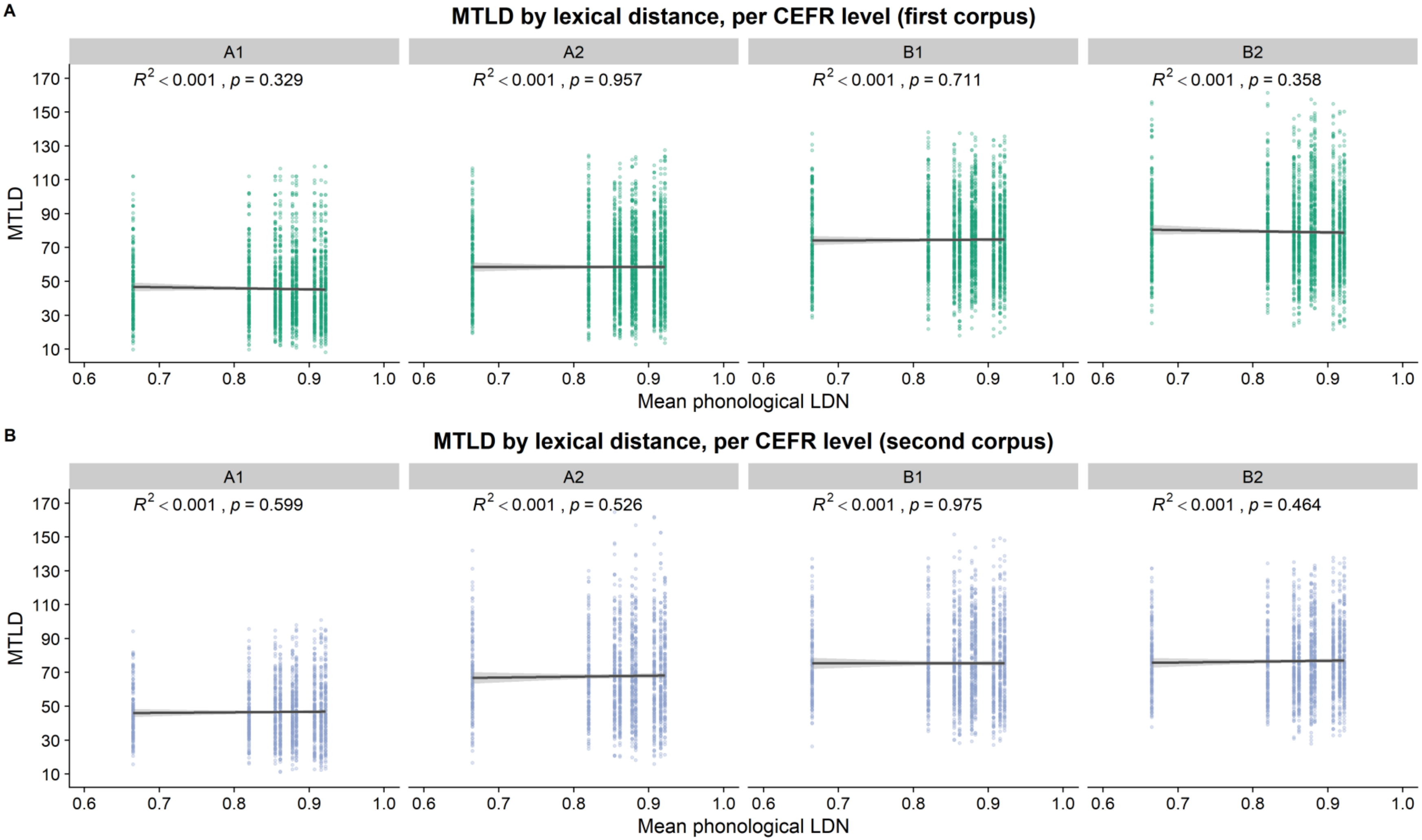


Figure 7. Scatterplots and linear models with lexical distance (LDN) as the predictor and lexical diversity (MTLD) as the response variable, per CEFR (L2 proficiency) level. Each model's R^2 and p appear in the corresponding panel. The grey bands around each line denote its 95% CI. Darker points denote a higher concentration of observations. The lexical distance range was 0.67–0.92. The MTLD range was 8.01–161.29 in the first corpus and 11.12–164.64 in the second. The number of observations at each CEFR level appears in Table 4 (range = 1,500–2,088).

Table 5. *Estimated marginal mean* (EMM) lexical diversity (MTLD) per L1 in each corpus, while controlling for *L2 proficiency* as a covariate, and for *task* and *learner* as random effects. There were 8,081 texts in the first corpus and 6,129 in the second.

	Lexical distance		MTLD (first corpus)		MTLD (second corpus)		Difference in rank [d]
L1	numeric	rank [a]	EMM (SE) [95% CI] [b]	rank [c]	EMM (SE) [95% CI] [b]	rank [c]	
German	.665	1	65.02 (1.38) [62.31, 67.73]	5	66.66 (1.54) [63.64, 69.67]	5	0
Italian	.820	2	66.58 (1.39) [63.86, 69.31]	2	65.82 (1.54) [62.80, 68.84]	6	-4
French	.855	3	64.95 (1.39) [62.23, 67.68]	6	67.92 (1.55) [64.89, 70.95]	3	3
Spanish	.862	4	61.37 (1.39) [58.64, 64.10]	9	62.61 (1.54) [59.60, 65.62]	9	0
Portuguese	.878	5	65.29 (1.38) [62.59, 67.99]	4	65.61 (1.52) [62.63, 68.59]	7	-3
Russian	.883	6	68.22 (1.38) [65.52, 70.93]	1	69.35 (1.54) [66.32, 72.37]	2	-1
Japanese	.907	7	64.61 (1.41) [61.85, 67.36]	7	67.43 (1.57) [64.36, 70.50]	4	3
Arabic	.916	8	61.93 (1.41) [59.16, 64.70]	8	64.43 (1.55) [61.41, 67.46]	8	0
Mandarin	.922	9	66.10 (1.38) [63.39, 68.81]	3	70.55 (1.55) [67.51, 73.58]	1	2

Note. The mean L2 proficiency level in both corpora was ~6.56 (on a scale of 1–12, corresponding to CEFR A1–B2).

[a] Ranked in *increasing* order of lexical distance (i.e., a rank of *1* denotes the smallest distance, and therefore the L1 that is lexically closest to English).

[b] The means are based on model estimates, so standard errors and confidence intervals are listed here (rather than SD).

[c] Ranked in *decreasing* order of mean MTLD (i.e., a rank of *1* denotes the highest MTLD, and therefore the L1 with the highest lexical diversity).

[d] This is equal to the MTLD rank of the L1 in the first corpus minus its rank in the second corpus.

Table 6 contains the main mixed-effects models of the study. The lack of significance of the lexical distance predictor and its interaction with L2 proficiency, together with their negligible effect sizes, indicate that lexical distance does not influence lexical diversity, either in general or at low L2 proficiency levels. Furthermore, the very small random effect of L1 further suggests that there is almost no difference in lexical diversity between speakers of different L1s, regardless of crosslinguistic lexical similarity.[12] Conversely, the significance and magnitude of the L2 proficiency predictor indicate that increased L2 proficiency predicts greater lexical diversity.

In addition, the large variance in lexical diversity between tasks, based on the associated random effect, indicates that there are substantial task effects. Specifically, the SD of between-task variability was 11.82 (i.e., the square root of the variance, equal to √139.66) in the first corpus and 11.14 (√124.01) in the second corpus. Since the expected increase of MTLD per EFCAMDAT level (the L2 proficiency measure) is 3.82 in the first corpus and 3.10 in the second, the SD of between-task variability is roughly equivalent to the expected increase of MTLD brought by 3.09 EFCAMDAT levels in the first corpus and 3.59 EFCAMDAT levels in the second. This, in turn, corresponds to a bit more than a whole CEFR level (1 CEFR level—e.g., A1—corresponds to 3 EFCAMDAT levels).

Finally, these findings are supported by additional models, presented in the supplementary information. Most notably, baseline models with no lexical distance led to similar results for *L2 proficiency* and *task* as the main models, and a comparison of the main models with the baseline models (based on AIC/BIC) provided support for the baseline models, which supports the finding that lexical distance does *not* predict lexical diversity, and does *not* interact with L2 proficiency. Furthermore, the findings were replicated in models that use binary distance from English (based on language family instead of LDN), and in models that use MATTR instead of MTLD as the measure of lexical diversity. These models all appear in the supplementary information.

[12] Though the magnitude of this random effect should be interpreted with caution, given the relatively small number of groups involved.

Table 6. Results of the mixed-effects linear regression models, with lexical diversity (MTLD) as the response variable. Under fixed effects, *lexical_distance* is the mean lexical distance between the L1 and English (0–1), and *L2_proficiency* is the EFCAMDAT level associated with each text (1–12). In the statistics, *std. B* and *std. 95% CI* provide information on the standardized coefficients, which were calculated by refitting the model on standardized data. Under random effects, σ^2 denotes the residual variance, τ_{00} denotes between-subjects (or groups) variance, *ICC* denotes the intraclass correlation coefficient, and *N* denotes the number of data points within each sampling unit. Finally, *observations* denotes the total number of texts in each sample, *Mar. [Marginal] R^2* denotes the proportion of the variance described by the fixed effects, and *Cond. [Conditional] R^2* denotes the proportion of the variance described by both the fixed and random effects.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	64.91	1.43	62.10 – 67.71	<0.001	0.01	-0.10 – 0.12	66.70	1.58	63.61 – 69.80	<0.001	0.00	-0.12 – 0.13
Lexical_distance	-2.35	10.25	-22.43 – 17.73	0.818	-0.01	-0.06 – 0.05	4.38	11.60	-18.36 – 27.12	0.706	0.01	-0.06 – 0.08
L2_proficiency	3.82	0.36	3.12 – 4.51	<0.001	0.50	0.41 – 0.59	3.10	0.39	2.34 – 3.86	<0.001	0.43	0.32 – 0.54
Lexical_distance * L2_proficiency	-0.11	0.88	-1.83 – 1.61	0.900	-0.00	-0.02 – 0.02	0.18	1.02	-1.81 – 2.17	0.858	0.00	-0.02 – 0.02
Random Effects												
σ^2	337.86						313.01					
τ_{00}	34.07_{Learner}						66.49_{Learner}					
	139.66_{Task}						124.01_{Task}					
	4.75_{L1}						6.04_{L1}					
ICC	0.35						0.39					
N	9_{L1}						9_{L1}					
	5385_{Learner}						4357_{Learner}					
	95_{Task}						71_{Task}					
Observations	8081						6129					
Mar. R^2 / Con. R^2	0.249 / 0.509						0.184 / 0.499					

[a] These ICC values can be interpreted as indicating that a medium portion of the total variance is explained by the grouping structure in the population (Hox et al., 2018).

3.4 Discussion

We examined the effects of *crosslinguistic lexical similarity*, *L2 proficiency*, and *task* on the lexical diversity of learners' L2 English writing. We used two matching sub-corpora with thousands of texts, written by speakers of nine typologically diverse L1s, in the A1–B2 range of CEFR L2 proficiency. We discuss our key findings below.

3.4.1 The effect of crosslinguistic lexical similarity on lexical diversity

Our results show that there is no effect of crosslinguistic lexical similarity on L2 lexical diversity, either in general or at early L2 proficiency levels. This was evident in the mixed models (Table 6), where lexical distance and the interaction between distance and L2 proficiency were non-significant and functionally zero (i.e., there was no distance effect and this did not change as learners' L2 proficiency changed). This was also evident in the estimated marginal means (Table 5), where there were only small differences between the L1s in lexical diversity, and where L1s that are lexically similar to English (e.g., German and French) sometimes had lower lexical diversity than L1s that are more lexically distant from English (e.g., Mandarin). Finally, this was also evident in the plots containing the association between lexical distance and lexical diversity across the CEFR levels (Figure 7), which also show that lexical distance is not a significant or substantial predictor of lexical diversity at any proficiency level.

These results suggest that the facilitative effect of lexical similarity on processing and learning that was found in prior studies does not translate to substantial differences in L2 lexical productions, at least when it comes to global measures such as lexical diversity (as opposed to the usage patterns of individual words), and when it comes to the present task-based settings.[13] This supports the findings of Crossley and McNamara (2011), on the lack of effect of L1 on lexical diversity in task-based settings, which they term *intergroup homogeneity*. The reason for this finding is likely that, in such settings, which learners will encounter in many educational contexts, lexical choices are driven primarily by task-based materials and communicative needs. Our findings, therefore, suggest that the effect of lexical similarity is

[13] We also considered and ruled out two other explanations. The first is that the number of cognates in the L1s is too low to influence lexical diversity, which would not explain why even highly similar L1s did not have higher lexical diversity than the distant L1s (as shown in Table 5). The second is that *false cognates* (or *false friends*)—words with similar form but different meanings across languages—may hinder L2 acquisition in L1s that are similar to the L2, but false cognates are generally much rarer than cognates (Ringbom, 2007), and this would contradict studies that found a facilitative effect of similarity on broad L2 proficiency (Schepens et al., 2020; Schepens, van der Slik, et al., 2013b; van der Slik, 2010).

limited, so even though it can facilitate processing, comprehension, and learning, learners ultimately cannot meet the communicative needs in L2 tasks without using extensive vocabulary, so they must acquire, learn to use, and use in practice necessary L2 vocabulary, regardless of its similarity to their L1.

From a practical perspective, these findings suggest that in language teaching and assessment, educators and assessors should generally expect learners to have similar lexical diversity, regardless of the lexical similarity between their L1 and the target L2, at least in certain task-based settings. This means, for example, that educators should generally not expect Mandarin speakers to have lower lexical diversity in their English essays than German speakers, even though German is more lexically similar to English.

Finally, a limitation of our study is that the placement of learners in specific proficiency levels in the EFCAMDAT might neutralize the potential L1 effect on lexical diversity, which could potentially appear in other samples. However, past studies on the EFCAMDAT found L1 effects across L2 proficiency levels for various linguistic phenomena, including clause subordination (Chen et al., 2021), relative clauses (Alexopoulou et al., 2015), clause-initial prepositional phrases (X. Jiang et al., 2014), grammatical morphemes (Murakami, 2016), capitalization (Shatz, 2019), and preference for certain punctuation marks and phrases (X. Jiang et al., 2014). Furthermore, also found similar L1 effects when comparing effects in the EFCAMDAT with those in other samples, such as the Cambridge Learner Corpus (Murakami & Alexopoulou, 2016). This suggests that L1 effects can occur in this sample, and that the lack of effect of crosslinguistic similarity on lexical diversity is likely due to a difference between crosslinguistic influence on lexical as opposed to functional content, the former playing a more important role in meeting communicative demands of tasks. Future research can confirm the generalizability of these findings in various ways, and most notably by replicating the analyses on other learner samples, or by examining a different target L2 (especially one that is not a lingua franca like English).

3.4.2 The effect of L2 proficiency on lexical diversity

As shown in the results (in Figures 5 and 6, and Tables 4 and 6), lexical diversity increased as learners' L2 proficiency increased, though it plateaued (i.e., slowed down) over time. There was some difference in this plateau between the two corpora, since in the second corpus MTLD plateaued consistently as L2 proficiency increased, while in the first corpus this association was more linear going from A1 to B1. Nevertheless, the mean MTLD per CEFR level was highly similar between the two corpora at all CEFR levels except for A2, and even in the first corpus, the increase in MTLD between the highest CEFR levels in the sample (B1–B2) was substantially smaller than between the other levels, both in terms of raw MTLD and in terms of percentage. In other words, it seems that the MTLD increase over L2 proficiency slows down as L2 proficiency increases, especially after the B1 level.

Our findings therefore support and extend past findings, especially given our use of a relatively broad sample and our focus on the association between L2 proficiency and lexical diversity in our analyses. Most notably, our findings support and extend Treffers-Daller et al. (2018), who also focused on the association between L2 proficiency and lexical diversity, and who examined lexical diversity at the B1–C2 range in the written essays of students taking the Pearson Test of English Academic. Specifically, they found similar mean values of MTLD as we found here: 70.14 at B1 (74.52 and 75.36 here), and 84.55 at B2 (79.20 and 76.62 here). Furthermore, similarly to us, they also identified a plateau (i.e., slowdown) in the increase of lexical diversity, so they were able to use MTLD to distinguish significantly only between texts at the B1 and C1 or C2 levels.

From a practical perspective, given the association between L2 proficiency and lexical diversity that was found here and in Treffers-Daller et al. (2018), including the large and fairly consistent within-level variance in lexical diversity across CEFR levels (Figure 5 and Table 4), it seems that MTLD can be used to distinguish between learners based on their L2 proficiency. However, our study suggests that MTLD can be most useful when used primarily on large-scale group data at the CEFR A1–B1 range, with lower confidence as learners' L2 proficiency increases, and preferably in conjunction with other measures of L2 proficiency.

3.4.3 The effect of task on lexical diversity

There were substantial task effects on lexical diversity, as shown by the between-task variance in Figure 5 and Table 6, which aligns with past findings (Alexopoulou et al., 2017; Michel, 2017; Michel et al., 2019). Our study further quantified the magnitude of these effects, and showed that task effects on lexical diversity (i.e., the standard deviation of between-task variation) can be of a magnitude equivalent to the increase in lexical diversity brought on average by a whole CEFR level.

From a theoretical perspective, this indicates that lexical diversity is strongly influenced by the functional goals and communicative needs associated with specific tasks, as suggested in §3.4.1. From a practical perspective, this highlights the importance of accounting for task effects when assessing lexical diversity.

3.4.4 Conclusions

In our study, lexical similarity between learners' L1 and their target L2 did *not* increase or otherwise influence their L2 lexical diversity, regardless of their L2 proficiency. This suggests that the facilitative effect of lexical similarity on processing and learning does not necessarily extend to L2 production, at least in the case of certain global measures, such as lexical diversity, and certain task-based settings, where lexical choices are driven primarily by communicative needs. In addition, lexical diversity initially increased as learners' L2 proficiency increased but then plateaued, and there were substantial task effects on lexical diversity. These findings are important to take into account when it comes to language teaching and assessment, as they help understand and predict the patterns that appear in learners' L2 lexical diversity.

4 Study 2: The influence of crosslinguistic similarity on L2 word choice

This chapter outlines a study that examines how similarity between L1-L2 words influences the usage patterns of the L2 words. As such, it is investigating a similar question as the previous study, but at a higher resolution, so to speak, by looking at the similarity between individual words (rather than languages as a whole), and by looking at the usage patterns of those words (rather than at the global lexical diversity of texts).

The abstract of the paper is as follows:

> We examined whether and how L1-L2 crosslinguistic formal lexical similarity influences L2 word choice. Our sample included two learner subcorpora, containing 8,500 and 6,390 English texts, written in an educational setting by speakers of diverse L1s, in the A1–B2 CEFR range of L2 proficiency. We quantified similarity based on phonological overlap between L1 words and their L2 (English) translations, and modeled the influence of similarity on the rate of use of the L2 words, while controlling for background factors, such as the baseline English frequency of words. This type of crosslinguistic similarity did not influence the choice of L2 words, regardless of learners' L2 proficiency. Conversely, there were strong task effects. This suggests that the influence of such similarity (which relates to cognancy) is constrained, and that communicative needs and task effects can override transfer, which raises questions regarding when and how else situational factors can influence transfer.

As with the previous study, because this chapter is written in the format of a paper, you will have already encountered some of the material here previously in the thesis, and especially in the previous study, example when it comes to the research background, methodology, and discussion. However, while there is some overlap in the material between this and the previous study, the two are distinctly different, and each represents a unique and independent contribution to the research literature.[14]

This work was done in collaboration with Dora Alexopoulou and Akira Murakami.

[14] This includes, for example the use of a different subset of the ASJP, the use of an additional lexical-distance dataset, the use of a different type of response variable (which does not have the same limitation as lexical diversity measures when it comes to short texts), the use of different (though similar) statistical models, and the use of different supporting analyses outside the main mixed models.

4.1 Introduction

[Due to space constraints—and to avoid repetition—I removed the opening of this introduction, which largely repeats the information from the thesis's introduction (§1.1 and §1.2) on transfer, crosslinguistic influence, crosslinguistic similarity, and cognate facilitation. Instead, we turn immediately to the discussion of the key study by Rabinovich et al.]

…One exception [to studies on word-choice transfer that did not investigate the effects of crosslinguistic similarity] is a study by Rabinovich et al. (2018), who showed that L1-L2 similarity can influence L2 word choice. Specifically, they investigated the spontaneous productions on a social media website—Reddit—of highly proficient (near-native) L2 English speakers of various Indo-European L1s. They focused on English words that were a part of a *synset*, which is a set of at least two synonyms that correspond to the same meaning. More specifically, they focused on synsets where the synonyms had at least two different etymological paths, and the synonyms themselves were fairly interchangeable.[15] They found clear evidence of a cognate facilitation effect, meaning that the speakers were more likely to use English words that are cognate with their L1.

However, there is also some evidence suggesting that the effect of crosslinguistic similarity may not always extend to productions in task-based settings, where there are strong task effects and constrained communicative constraints. Specifically, Crossley and McNamara (2011) found that L2 texts written by speakers with different L1s had similar scores on several global lexical measures, such as lexical diversity and polysemy, despite different levels of similarity between their L1s and the target L2. This is based on an examination of 599 L2 English texts in the *International Corpus of Learner English*, written by Czech, Finnish, German, and Spanish speakers, who are "high intermediate to advanced" L2 English speakers (p. 274), and who wrote the texts as a response to one of few prompts for argumentative essays. This finding is unexpected given the L1 effects that were found in other studies, and casts uncertainty regarding whether the associated L1 effects play a role in learners' L2 word choice in task-based settings.

To summarize, there is clear evidence for a facilitative effect of L1-L2 lexical similarity on L2 processing, comprehension, and learning, particularly at the early stages of L2 acquisition (e.g., Ringbom, 2007; Schepens et al., 2020), and there is also evidence showing that learners'

[15] They operationalized interchangeability as meaning that no synonym within the synset accounted for more 90% of the occurrences of the synset in their corpus; see the supplementary information for a brief discussion of this operationalization.

L1 can influence their L2 word choice (e.g., Jarvis et al., 2012). However, evidence regarding the influence of crosslinguistic similarity, in particular, on L2 word choice is limited and less clear, especially in task-based settings. Based on this, we ask the following research questions:

1. Does increased similarity between L1 words and their L2 translations lead to increased use of the L2 words in task-based English-as-a-foreign language (EFL) settings?
2. If there is an effect of crosslinguistic lexical similarity in such task-based settings, is it moderated by learners' L2 proficiency?

Investigating these questions will help reconcile the different findings on the topic, and will shed light on remaining gaps in knowledge. Notably, it will help determine whether the effect identified by Rabinovich et al. (2018) extends to task-based settings, whether findings regarding word-choice transfer in task-based settings are likely attributable to some degree to crosslinguistic similarity, and whether the lack of L1 effect (i.e., *intergroup homogeneity*) found by Crossley and McNamara is a simply feature of the global lexical measure that they used and/or their sample. In addition, the focus on an educational EFL setting will shed light on the influence of crosslinguistic similarity in this type of common environment, where, as we will see, there are often strong task effects on word choices.

In this study, we will examine how similarity between L2 English words and their L1 translations influences the usage rates of the L2 words. For example, we want to see if an Italian learner of English will be more likely to use the word "lemon" than a French speaker under the same circumstances, because the Italian word for "lemon" ("limone") sounds more similar to the English word than the French one ("citron") does. If similarity does matter, then, we expect that, where possible, learners will prefer to use similar words, because they are easier to learn and retrieve.

To investigate this, we will construct two crosslinguistic word lists of L1-L2 pairs (e.g., lemon-citron), and calculate the similarity for each pair, based on their phonological overlap. Then, we will build mixed-effects models to see if L1-L2 similarity influences word choice in an EFL learner corpus that covers a wide range of tasks, L2 proficiency levels, and L1s, while controlling for relevant factors such as the baseline frequency of the L2 words.

4.2 Methodology

Below, we first present the methods we used to calculate lexical distance (§4.2.1) and baseline English word frequency (§4.2.2). Then, we introduce our learner sample (§4.2.3), and explain how we calculated L2 usage rates (§4.2.4). Finally, we present the models that we used to analyze the data (§1.1.1).

All the study's data and code are available at the following *Open Science Framework* (OSF) repository: https://doi.org/10.17605/OSF.IO/5EUA8

4.2.1 Lexical distance

4.2.1.1 Distance datasets

In the present research, we quantify crosslinguistic lexical similarity based on the lexical distance between L1 words and their L2 translations, where increased distance denotes lower similarity.[16] To do this, we use two datasets, which contain lists of corresponding words in different languages, as outlined briefly below. For more information on these datasets and their processing, see the comprehensive "Lexical-distance datasets information" document in the OSF repository [included in the thesis as Appendix C].

The first lexical-distance dataset is the *Automated Similarity Judgment Program* (ASJP) (Wichmann et al., 2018). It contains Swadesh lists, which are often used by researchers to calculate the lexical distance between languages (e.g., Schepens et al., 2013), and which contain words representing various concepts, such as *hear*, *water*, *full*, *one*, and *dog* (Swadesh, 1950; Wichmann et al., 2018).

The Swadesh lists in the ASJP focus on a subset of 40 concepts, and to control for variation in the completeness of the Swadesh lists across languages, we included in our analysis only the 38 concepts that are shared by all the languages in our sample. These languages, which are based on the ones available in the learner sample that is outlined in §4.2.3, are: Arabic, French, German, Italian, Japanese, Mandarin, Portuguese, Russian, and Spanish as L1s, and English as the target L2. In addition, we focused on single-word entries, in line with most prior research and to avoid potential confounds, and so we included only entries that do not contain a multi-word phrase in any of the L1s or English. Accordingly, the final Swadesh-based sample contains 225 entries, with 25 entries for each of the 9 L1s, where each entry is a row containing an English word together with all its L1 counterparts in a specific L1.

However, an important caveat about the ASJP Swadesh lists is that they were collected primarily with the goal of comparing crosslinguistic distance at the language level, rather than

[16] We use the term *lexical distance* to distinguish it from other types of linguistic distances, as explained in the supplementary information.

at the word level. As such, they contain a small number of words per L1, as shown above, and these words are all relatively high-frequency words, as will be shown in the section on baseline word frequency. This does not prohibit the use of the Swadesh lists in our analyses, but it does mean that they are not, by themselves, sufficient in order to investigate our research questions.

As such, we extended our analyses by also using a second lexical-distance dataset—the *Intercontinental Dictionary Series* (IDS)—which contains parallel dictionaries in various languages (Key & Comrie, 2015). Similarly to the Swadesh lists, this dataset also contains a standardized list of words and their corresponding counterparts in various languages. But, the parallel dictionaries contain substantially more words per language than the Swadesh lists (~1300 general word meanings compared to ~40), which include both high- and low-frequency words (as shown in the section on baseline word frequency). However, a disadvantage of the parallel dictionaries is that they contain data only for French, German, Italian, Portuguese, and Spanish in the present sample. As such, they serve complement to the Swadesh lists, but do not replace them. Furthermore, it is beneficial to use multiple sources of lexical-distance data, as there can be various idiosyncratic issues with the transcriptions used in any given dataset; for more information on this in the context of the Swadesh lists, see the overview of these lists in Appendix C. Nevertheless, since the parallel dictionaries contain substantially more data than the Swadesh lists—as they cover more words by more than an order of magnitude, even taking into account the smaller number of L1s—the parallel dictionaries should be viewed as the main source of lexical-distance data in the present study.

As with the Swadesh lists, we included only single-word entries in our analysis of this dataset. Furthermore, we removed from the parallel dictionaries a small number of words (22) that also appeared in the Swadesh lists, so that each dataset contained a unique lexical sample. Accordingly, the final parallel-dictionaries sample contains 5,515 entries, with 1,103 entries for each of the 5 L1s, where each entry is a row containing an English word and all its corresponding L1 counterparts in a single L1.

4.2.1.2 Calculating lexical distance

The lexical-distance measure that we used is *Levenshtein distance normalized* (LDN).

Intuitively, LDN generally represents the degree of phonological or orthographic overlap between two words. It is calculated by taking the minimum number of character substitutions, additions, and deletions that are needed to transform one string to another (i.e., the *Levenshtein distance*), and dividing it by the length of the longer string, to account for variations in word length. For example, in the case of the word *knee*, the English-German pair /ni/-/kni/ has an LDN of 0.33, since there is 1 character transformation (a /k/ is inserted or deleted), and the length of the longer string is 3. By contrast, the LDN for the corresponding English-Japanese pair /ni/-/hiza/ is greater (0.75), since there is less overlap, so more transformations are needed.

In the present research, we first calculated lexical distance between each L1 entry and its corresponding L2 (English) entry, based on their phonological (IPA) transcription. When there were multiple L1 synonyms available (e.g., "soil" in French—*sol* and *terre*), we used the distance from the closest synonym, as our goal was to identify cases where the L2 word is closely similar to an L1 word (and is likely to be cognate with it).

We used LDN for several reasons. First, this measure can be calculated in an automated, objective, and replicable manner for a large number of words from different languages (Schepens et al., 2012). Second, it is the most conventional measure that is used for this purpose, and, as shown in detail in the supplementary information (under "Validation of Levenshtein distance"), it has been extensively validated in past research in various fields, including typology, psycholinguistics, and SLA. This validation includes comparisons with other measures of language distance, such as expert cognancy judgments from historical linguistics and perceived language distance from psycholinguistics, which showed strong correlations and convergent validity between LDN and the other measures (e.g., Beijering et al., 2008; Schepens et al., 2012). Furthermore, LDN has been used extensively by many SLA researchers (e.g., De Wilde et al., 2020) in a similar manner as we use it here, to quantify crosslinguistic similarity between individual words—often to distinguish cognates from non-cognates when investigating the effects of cognate facilitation—and it has been shown to be a robust predictor of many L2 outcomes, including word recognition (Carrasco-Ortiz et al., 2021), word retrieval (Sadat et al., 2016), and word processing speed and accuracy (Casaponsa et al., 2015).[17]

4.2.1.3 Limitations of LDN

LDN is limited in several key ways.

[17] Some studies used non-normalized LD or orthographic LD, as discussed in the supplementary information.

First, it treats all character transformations as equal. For example, this means that the English word “fish” /fɪʃ/ has an equal and maximal LDN of 1 from both the corresponding Spanish word (“pez” /pes/) and the Hebrew one (“דג” /dag/), even though the English word is closer phonologically and etymologically to the Spanish word than to the Hebrew one, and could be considered a cognate of the first but not the second.

To partially address this issue, we replicated our analyses using *feature edit distance* (or *phonological edit distance*), which assigns different costs to the transformation of different phonological units. The results of these models replicated our results when using LDN as the measure of distance, as shown and explained in detail in the supplementary information. Briefly, this distance, which has less validation and standardization than LDN, attempts to account for the phonological similarity across segmental units, by assigning different costs to the transformation of different units, based on their phonological features. For example, in the case of “fish” considered above, substituting /ʃ/ with /z/ generally incurs a lower cost than substituting /ʃ/ with /g/, since /ʃ/ and /z/ share more phonological features (e.g., being coronal), so they are more similar to each other from a phonological perspective.

Another limitation of our use of LDN as a measure of lexical distance is that it only looks at one aspect of formal similarity across words (phonological overlap). However, other factors, including both formal ones, such as orthographic depth, and non-formal ones, such as semantic and pragmatic similarities, may also affect crosslinguistic influence. For example, it may be that there is an interaction between orthographic depth and the effects of phonological distance, or that the use of a different script across L1s from different language families moderates the effects of phonological similarity.

Some past studies (e.g., Sadat et al., 2016) found a facilitative effect of formal similarity even without considering these factors, so we expect to be able to do the same. In addition, we further partially addressed this limitation in our analyses, by using mixed-effects models to control for some of these potential effects through random effects for *word* and *L1*. However, future analyses may still benefit from assessing the role of these factors directly.

Finally, note that LDN does not assess *cognancy* directly, which we define here in the psycholinguistic sense, of words that have similar meaning and pronunciation/spelling across languages. Rather, it only quantifies the formal similarity between words that are generally similar in terms of meaning. Most notably, this means that there are cases where a large distance does not indicate lack of cognancy, as in the “fish” example above. Nevertheless, as noted above, LDN is strongly correlated with cognancy (e.g., Schepens et al., 2012), and has been used to estimate cognancy directly in SLA studies that then used it to successfully predict L2 outcomes (De Wilde et al., 2020; Sadat et al., 2016), so we expect to be a reasonable approximation in the context of the present large-scale analyses.[18]

It is important to keep these limitations in mind when interpreting the findings of the study. Nevertheless, as noted in the previous sub-section, and as explained in the supplementary information (under "Validation of Levenshtein distance"), this distance has been extensively validated through research in various fields. This validation includes, most notably, strong correlations with other measures of distance, such as expert cognancy judgments and perceived language distance (Beijering et al., 2008; Schepens et al., 2012), and the use of this measure in SLA to successfully predict many L2 outcomes at the word level—including in the context of the cognate facilitation effect—such as word recognition and word production, in a similar manner as in the present study (Carrasco-Ortiz et al., 2021; Sadat et al., 2016). As such, we believe that the use of LDN is reasonable in the present study. Most importantly, even if it will be unable to perfectly capture *all* of the effects of crosslinguistic similarity, it should be able to successfully capture some of them, as it did in many past SLA studies.

4.2.1.4 Lexical distances

Figure 8 and Table 7 contain information about the the lexical distances between the L1s in the sample and English. The distances of all word pairs are available in the data files in the OSF repository.

This figure and table show that the words in the datasets cover the full range of distances from English (0‒1), though most words tend to be highly dissimilar (with an LDN near 1). This suggests that learners may generally have limited opportunities to benefit from facilitative effects of crosslinguistic lexical similarity, although there was nevertheless a sufficient range of distances in our sample that the estimates of its effects were precise in our models, as shown in the results section.

In addition, the distances are largely aligned with those based on general language classification. Specifically, the Germanic and Romance L1s are the closest to English, and the

[18] We used the phonological distance directly, rather than using it to estimate cognancy, because there is currently no standardized and well-validated way to determine cognancy based on distance.

Indo-European L1s are closer to English than the non-Indo-European L1s, except that Japanese is shown as being closer to English than Russian. However, because the lexical-distance datasets were modified through the removal of multi-word entries, the mean similarity between each L1 and English here should *not* be interpreted as the mean similarity between that L1 and English. Indeed, as shown in the supplementary information (under "Validation of Levenshtein distance"), when the unmodified wordlists are used, meaning that multi-word entries remain in the sample, Japanese and Russian switch positions as expected, and consequently, all the Indo-European L1s are closer to English than the non-Indo-European L1s. Nevertheless, this is not important for our analyses, since we focus on the similarity and use of individual words, rather than on similarity at the language level and on global measures of word use (e.g., lexical diversity).

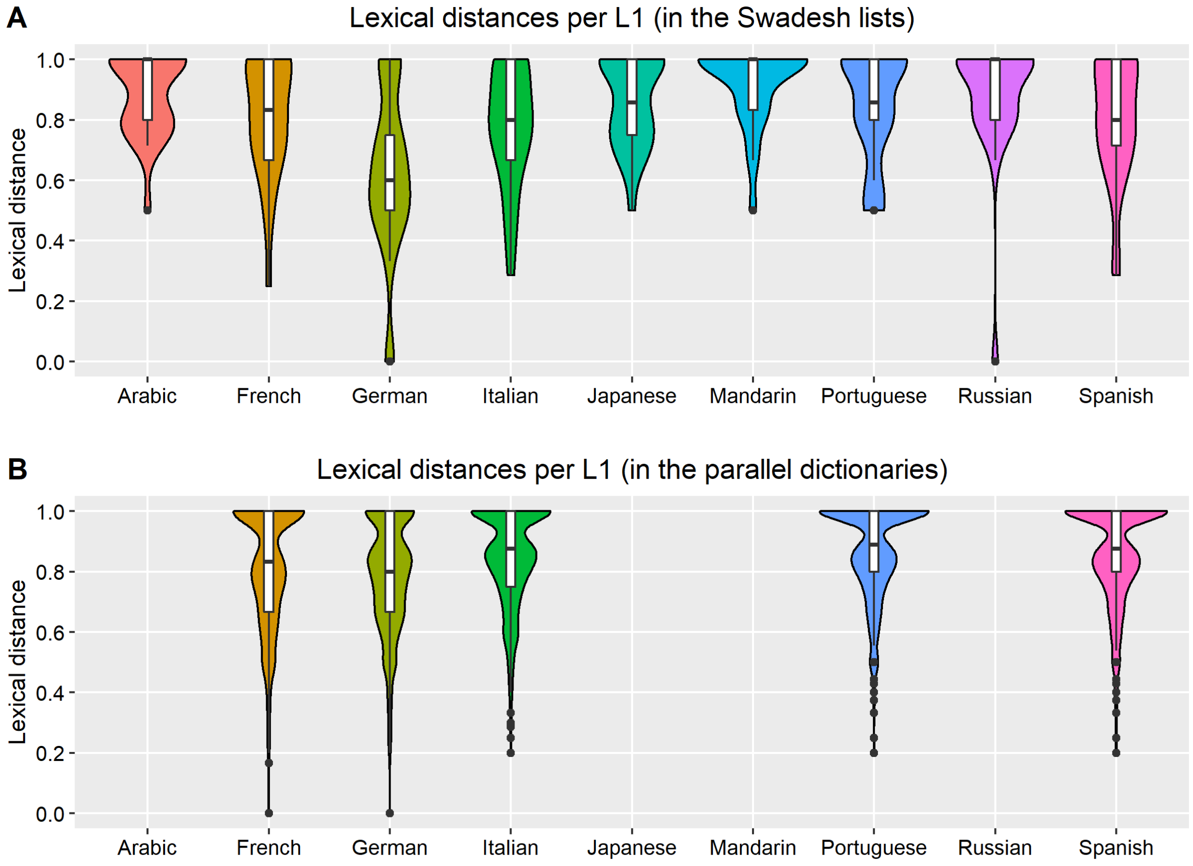


Figure 8. Lexical distance between L1 words English, per L1 in each dataset. The distance is equal to the phonological LDN between L1 words and their most lexically similar English counterpart. Within the boxplots, the line inside the box indicates the median, the lower and upper hinges indicate the 1st and 3rd quartiles, the whiskers indicate 1.5 interquartile ranges (IQR) past the hinges, and the dots indicate outliers beyond that. The violin plots indicate an estimate of the probability density of lexical distance for each L1, which can be viewed as the likelihood that a word in each L1 will have a certain lexical distance, where increased width indicates greater likelihood. Data is based on 25 words per L1 in the Swadesh lists and 1,103 words per L1 in the parallel dictionaries. Note the similarity in lexical-distance patterns between Portuguese and Japanese, despite Portuguese being an Indo-European Romance language, and Japanese being non-Indo-European. However, as noted in the preceding discussion, this should not be interpreted as the true distance (at the language level) between these L1s and English, due to the systematic removal of multi-word entries from the sample.

Table 7. Statistics about the lexical distances between the L1s and English in each dataset. L1s are arranged in order of increasing mean lexical distance in the Swadesh lists.

	Swadesh lists					Parallel dictionaries				
L1	mean	SD	median	IQR	range	mean	SD	median	IQR	range
German	.622	.27	0.60	0.50-0.75	0.00-1.00	.785	.18	0.80	0.67-1.00	0.00-1.00
Italian	.776	.20	0.80	0.67-1.00	0.29-1.00	.847	.16	0.88	0.75-1.00	0.20-1.00
Spanish	.808	.21	0.80	0.71-1.00	0.29-1.00	.860	.16	0.88	0.80-1.00	0.20-1.00
French	.813	.20	0.83	0.67-1.00	0.25-1.00	.814	.20	0.83	0.67-1.00	0.00-1.00
Portuguese	.848	.18	0.86	0.80-1.00	0.50-1.00	.873	.15	0.89	0.80-1.00	0.20-1.00
Japanese	.864	.15	0.86	0.75-1.00	0.50-1.00	-	-	-	-	-
Russian	.881	.21	1.00	0.80-1.00	0.00-1.00	-	-	-	-	-
Arabic	.887	.14	1.00	0.80-1.00	0.50-1.00	-	-	-	-	-
Mandarin	.924	.13	1.00	0.83-1.00	0.50-1.00	-	-	-	-	-

Note. The distance here is the phonological LDN from the closest synonym, calculated for the single-word entries in each dataset. There were 225 entries in the Swadesh lists (i.e., rows with an English word and all its corresponding counterparts in a certain L1), with 25 entries for each of the 9 L1s in the dataset. There were 5,515 entries in the parallel dictionaries, with 1,103 for each of the 5 L1s.

4.2.2 Baseline word frequency

Baseline word frequency is how often an English word is used in English in general, which we needed to control for since it can influence the outcome that we will be looking at—the usage rate of individual L2 words in our sample. The "Baseline frequency information" document in the OSF repository [included in the thesis as appendix C] contains detailed information regarding how we calculated this frequency, and explains the rationale behind the process. To summarize, we used the *wordfreq* library in Python (Speer et al., 2018), which curates frequency information from a number of diverse and large-scale sources, including books, subtitles, news, and social media. The specific frequency measure that we used from the library is *Zipf frequency* (developed by van Heuven et al., 2014), which is the base-10 logarithm of the number of times a word appears per billion words. Accordingly, "A word with Zipf value 6 appears once per thousand words, for example, and a word with Zipf value 3 appears once per million words" (Speer et al., 2018).

Figure 9 shows the distribution of the frequencies of the English words that were included in our lexical-similarity datasets. The frequencies of all words are available in the data files in the OSF repository. The mean Zipf frequency in the Swadesh lists was 5.24 (SD = 0.72, median = 5.14, range = 4.15–7.11), and the mean Zipf frequency in the parallel dictionaries was 4.35 (SD = 0.83, median = 4.32, range = 1.87–7.41). As such, both datasets included a wide range of words with different frequencies, though this range was greater in the parallel dictionaries.

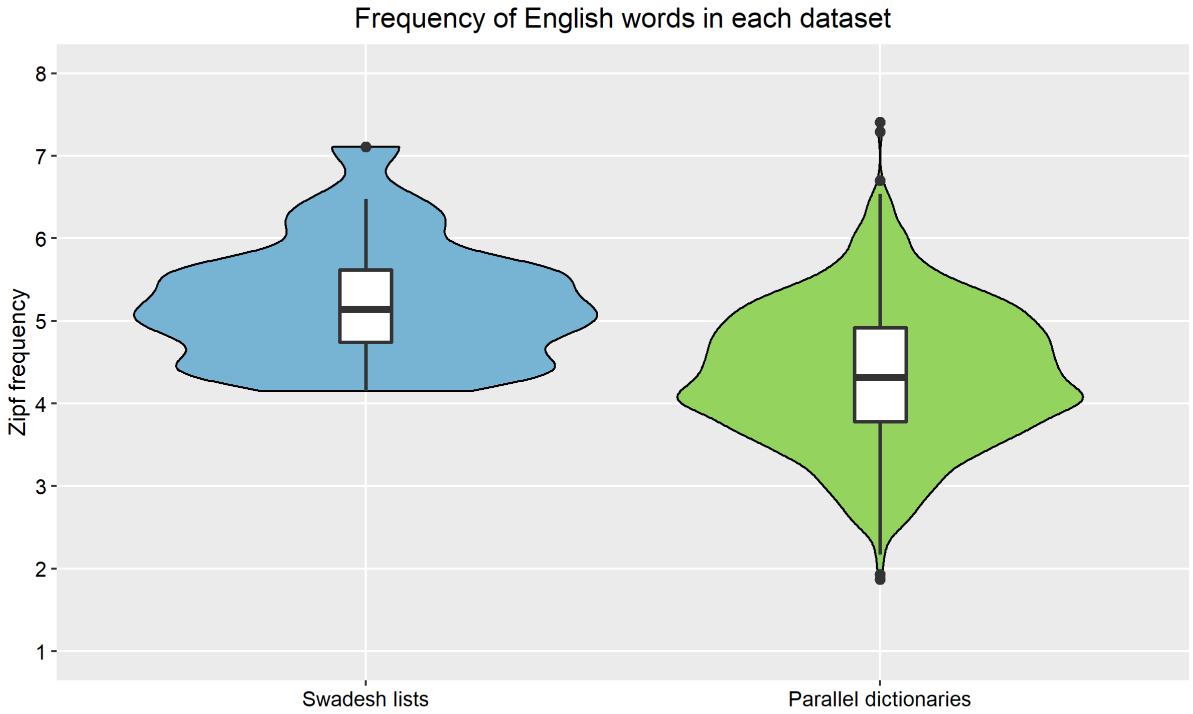


Figure 9. The baseline (Zipf) frequency of the English words included in each lexical-distance dataset. Within the boxplots, the line inside the box indicates the median, the lower and upper hinges indicate the 1st and 3rd quartiles, the whiskers indicate 1.5 IQRs past the hinges, and the dots indicate outliers. The violin plots indicate an estimate of the probability density of the frequency of English words. Data is based on 25 English words in the Swadesh lists and 1,103 words in the parallel dictionaries.

4.2.3 Learner sample

[I removed this section, since it duplicates the corresponding section in study 1, as both studies use the same sample. The only unique thing noted in this section is that, in the parallel-based samples, which contain only 5 L1s (compared to 9 in the Swadesh lists), there are 4,747 texts in the first corpus (55.85% of the 8,500 used with the Swadesh lists), and 3,550 in the second (55.56% of 6,390).]

4.2.4 Word usage

To assess learners' use of L2 vocabulary, we calculated the number of times each English word in the lexical-distance datasets appears in any given text in the learner sample. We did this separately for each cross of one of the lexical-distance datasets with one of the EFCAMDAT subcorpora, as shown in Table 8.[19]

Table 8. The four final samples, each representing a cross between a lexical-distance dataset and a subcorpus. *Observations* equal the number of *words per L1* in a lexical-distance dataset times the number of *texts* available in the subcorpus.

Distance dataset	Subcorpus	L1s	Words per L1	Texts [a]	Observations
Swadesh lists	first	9	25	8,500	212,500
Swadesh lists	second	9	25	6,390	159,750
Parallel dictionaries	first	5	1,103	4,747	5,235,941
Parallel dictionaries	second	5	1,103	3,550	3,915,650

[a] The number of texts available for the parallel-dictionaries samples reflects them containing data for 5 out of 9 L1s that we examine.

Statistics about the counts of target words appear in in Table 9. For more information on the raw response variable, see the section on "Correlations of distance, frequency, and word use" in the supplementary information.

Broadly, the data can be characterized as having a high proportion of zeros and a right skew, which means that there were many cases where a target was not used in a text, and a small number of cases where a target word was used in a text multiple times. This distribution is common for count data, and is expected given the diverse range of tasks and words in our sample, including the spectrum of low- and high-frequency words.[20] This distribution should

[19] We calculated counts based on a spelling-corrected version of each text, which comes built-in as part of the EFCAMDAT Cleaned Subcorpus, and which was generated using the *autocorrect* library (McCallum, 2019) in Python, since we are interested in how often learners attempt to use target words, and misspellings could obscure those patterns. Nevertheless, this does not appear to make a practical difference to our analyses, as the correlations between the corrected and uncorrected counts were extremely high (*Pearson's r* = .9954–.9998 for all datasets, with p < .001 in all cases, and the *95% CIs* falling no more than .0001 from the estimates. *Spearman's ρ* had similar values, from .9918–.9982, all with p < .001).

[20] This means that most words are not used in most texts, and that some words are also not used in any of the texts, which is expected, given that we include specialized "high level" (i.e., low frequency) words in our sample. However, the inclusion of such words does not pose an issue for the models, as indicated by the model diagnostics that are discussed later, and as later indicated by the precise coefficient estimates for our predictors. In addition, note that removing such words from our sample would bias the results. To illustrate this with an example, consider a simple situation, where we compare, among German learners, the rate of use of two English words, that have equally low baseline frequency. One of the words is distant from the corresponding German word, whereas the other is similar to it. In this case, if there is indeed a facilitative effect of similarity, then we would expect that the distant word will not be used by learners (because it has low baseline frequency), but we would expect the similar word to be used in spite of the low baseline frequency (because of the facilitative effect). However, if we remove

not be interpreted as being indicative of overdispersion or zero-inflation, since those are features of a model rather than the response variable (Hartig, 2021a). Indeed, the assumption checking (in the "Model diagnostics" section of the supplementary information) show that the models do are not overdispersed or zero-inflated; rather, some actually have underdispersion, though as shown in the aforementioned section, this does not substantially influence our results. Also, as noted in the next section, we used Poisson models in our analyses, since they are designed for dealing with this type of count data, and due to the large size of the samples, there was a sufficient number of "positive" observations (i.e., with a count > 0) that the models were able to converge properly.

the distant word from our analysis because it is not used at all, then we would be obscuring the effects of similarity by comparison. Essentially, the fact that a word is not used at all by learners is important to our analyses, as it allows us to more accurately determine the effects of distance. This was a simple example, meant to clearly and intuitively illustrate the associated issue, but the same principle applies to the more complex analyses that we used in practice.

Table 9. Statistics about the distribution of the count data that was used in the models (i.e., the number of times a word appeared in a text). The specific statistics are given either for all observations (*total*), or for observations where the count was greater than zero (*count>0*).

Dataset	Subcorpus	$N_{(total)}$	$N_{(count>0)}$	$Prop._{(count>0)}$ [a]	$Mean_{(total)}$	$SD_{(total)}$	$Mean_{(count>0)}$	$SD_{(count>0)}$	Max
Swadesh	first	212,500	13,049	0.061	0.174	0.968	2.832	2.782	24
Swadesh	second	159,750	9,819	0.061	0.188	1.104	3.063	3.323	26
Parallel	first	5,235,941	59,566	0.011	0.016	0.183	1.417	0.973	19
Parallel	second	3,915,650	47,072	0.012	0.017	0.196	1.452	1.058	15

Note. The difference in distributions between the parallel dictionaries and Swadesh lists could be attributed, at least in part, to the parallel dictionaries having containing some lower-frequency words. Specifically, the mean Zipf frequency in the Swadesh lists was 5.24 (SD = 0.72, median = 5.14, range = 4.15–7.11), while the mean Zipf frequency in the parallel dictionaries was 4.37 (SD = 0.84, median = 4.35, range = 1.87–7.41).

[a] This represents the proportion of entries with a count greater than 0, out of all entries in the sample.

4.2.5 Data analysis

We built generalized linear mixed-effects models (GLMM), separately for each combination of corpus and lexical-distance dataset (e.g., Swadesh lists in the first corpus). Specifically, we built *Poisson* models (with the canonical *log* link), due to the use of count data in the response variable (Hox et al., 2018; Winter, 2019). The structure of the models was as follows:

1. **Response variable**: *rate of usage* of the target English word. This is based on the number of times the target English word appears in a text, which is *offset* by the total number of words in the text,[21] to control for different texts having a different total number of words (Hox et al., 2018; Winter, 2019).
2. **Predictors**:
   a. *Lexical distance* (of individual L1-L2 word pairs), based on the phonological LDN between the English word and its closest synonym in the L1 of the learner who wrote the text.
   b. *L2 proficiency,* based on EFCAMDAT proficiency level (1–12, corresponding to CEFR A1–B2) of the learner at the time they wrote the text (each lesson/task is classified under a certain proficiency level). This predictor was added as both a main effect and as an interaction term with *lexical distance,* to see whether the effects of L2 proficiency moderate those of lexical distance.
   c. *Word frequency* of each English word (based on its baseline frequency in the English language), to control for this factor when considering the word's rate of usage in the L2 texts.
3. **Random effects** (random intercepts unless noted otherwise)[22]:
   a. *Learner,* to control for learners who had more than one text in the sample.[23]
   b. *L1,* with random slopes for *lexical distance,* to control for any other effects from the learners' L1 and their associated (e.g., cultural) background.
   c. *Task,* to control for all the aspects of each writing task that can influence word choice, such as its prompt, with the exception of the task's associated L2 proficiency level, which we control for using the relevant predictor. Note that this approach

[21] The total number of words in each text is based on the *wordcount* variable that is available in the EFCAMDAT Cleaned Subcorpus (Shatz, 2020).

[22] As discussed in the supplementary information, additional random intercepts and slopes were considered but not included in the models due to convergence issues, though this does not appear to substantially influence our findings.

[23] Most learners only had a single text in the sample (the mean number of texts per learner was 1.36 in the first corpus and 1.41 in the second). See the "Sample information" document in the OSF repository for more details.

accounts for all aspects of task effects in aggregate, and does not disentangle the different aspects.[24] In addition, note that the use of mixed-effects models allows us to assess such task effects despite the fact that each task is associated with only a single proficiency level (Hox et al., 2018; Winter, 2019), and this type of mixed-effects structure—where each group in a random grouping variable always takes the same potentially unique value along a continuous predictor—is conventional in both corpus linguistics (e.g., Levshina, 2018) and psycholinguistics (e.g., Baayen et al., 2007; Vandenberghe et al., 2021).[25]

d. *Word*, to control for word-level random effects, in a similar manner as discussed above for *task*.

e. *Task:Word*, to control for the effects of the interaction between *task* and *word*, and particularly cases where a certain task is more likely to prompt the use of a certain word.

Before building the models, we scaled the distance predictor by a factor of 10, so that it is on a scale of 0–10 instead of 0–1, to facilitate model convergence by putting this predictor on a similar scale as the other predictors (L2 proficiency: 1–12, frequency: ~1–7.5). We also centered the three predictors, to facilitate convergence of the models and reduce potential collinearity between predictors and the interaction term.

After building the models, we exponentiated the coefficient estimates to derive an *incidence rate ratio* (IRR), in order to facilitate the interpretation of the results, and the associated *standard errors* (SEs) were scaled accordingly (Hox et al., 2018; Sedgwick, 2010). The IRR can be interpreted as the expected change in the rate of the response variable as a factor of a 1-unit increase in the predictor. For example, an IRR of 2 means that a 1-unit increase in the predictor doubles the rate of response (i.e., doubles the rate of use of the target word), while an IRR of 0.5 means that a 1-unit increase in the predictor halves it. An IRR of 1 corresponds to a coefficient estimate (B) of 0, as there is no expected change in the response variable as a result of a change in the predictor. For more details regarding IRR, see the supplementary information.

[24] As such, this operationalization of task is distinct from most notions of *task* within task-based learning and teaching approaches, and we make no claim regarding the impact of any specific aspect of tasks, such as their genre or cognitive complexity (Alexopoulou et al., 2017).

[25] Specifically, Levshina (2018) used *website* as a random effect, and website *formality* (mean word length) as a predictor. Baayen et al. (2007) used *item* as a random effect, and item *frequency* as a predictor. Vandenberghe et al. (2021) used *participant* as a random effect, and participant *vocabulary size* as a predictor.

In addition, we checked the statistical assumptions of the models. The relevant diagnostics appear in the supplementary information, and indicate that there are no substantial issues with the models.

Finally, we also compared these models with baseline models, which did not include lexical distance as a predictor, to determine whether the inclusion of lexical distance improves the models' predictive power.

4.3 Results

Figure 10 contains plots showing the basic association between distance and the rate of use of words in the datasets, compared to their baseline frequency in English. For the associated statistics, see "Frequency-ratio descriptive statistics" in the supplementary information.

If there is facilitative effect of crosslinguistic similarity, then we would expect words with a lower lexical distance to have a higher frequency ratio; this would indicate that when a word in a learner's L1 is similar to the corresponding L2 English word, then they are more likely to use it, which would mean that they use it more often in their writing than the word is used in baseline English. However, such an effect is not clearly visible in the plots, where the frequency ratio seems independent of lexical distance. Nevertheless, since this analysis is fairly limited (e.g., it does not control for potential task effects), we do not rule out the presence of this facilitative effect based on it, and instead move on to the more comprehensive mixed-effects models.

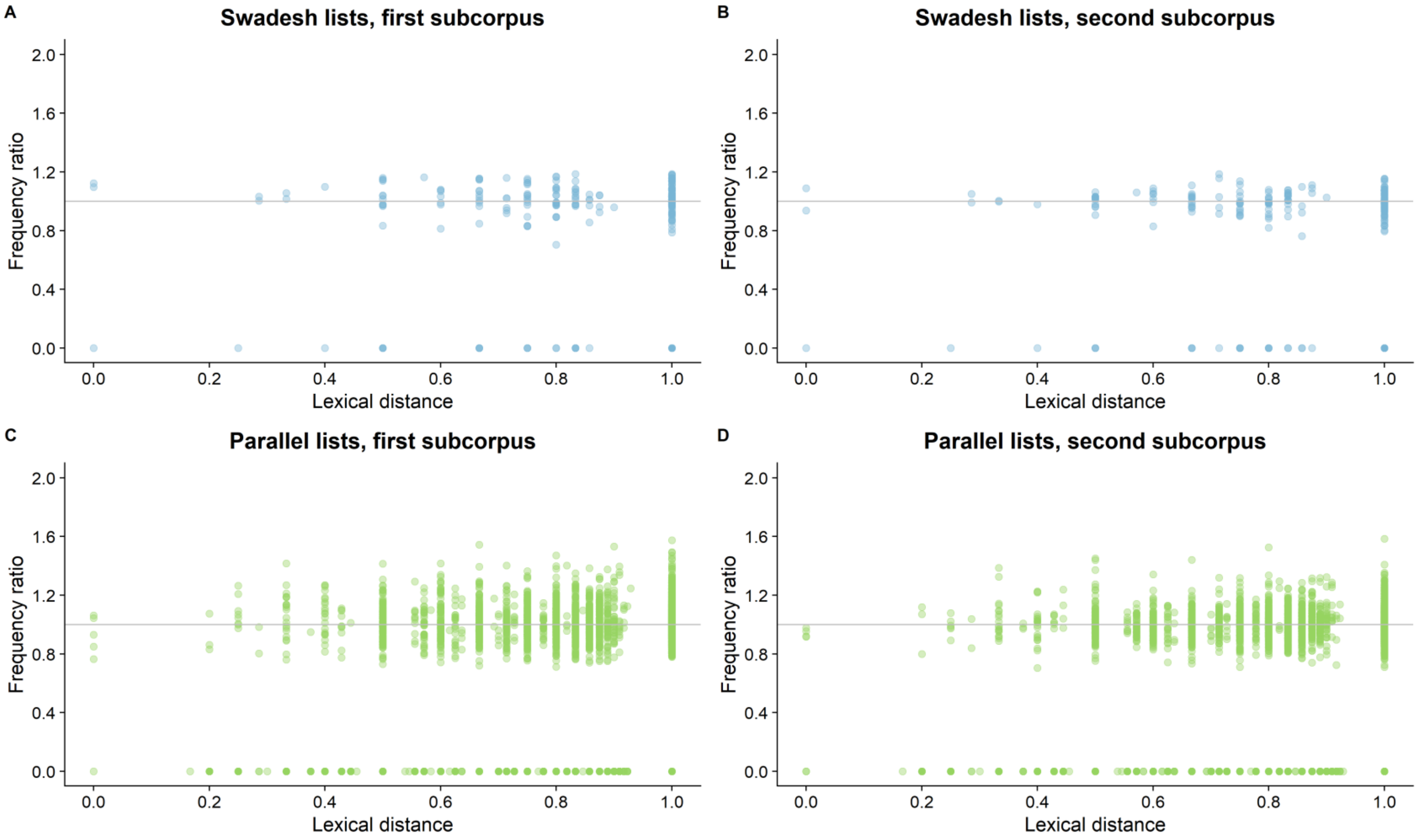


Figure 10. The *lexical distance* of words and their *frequency ratio* (i.e., their frequency in the sample divided by their baseline frequency in English, using the frequency measure described in §4.2.2). Accordingly, a ratio =1 (grey line) indicates that a word is used in an equal rate in the learner sample and in baseline English, whereas a ratio >1 indicates that the word is used more frequently in the sample, and a ratio <1 indicates the opposite. Words that did not appear in the sample were assigned a Zipf frequency of 0, in line with Speer (2020), and consequently have a frequency ratio of 0 here. Each point is a combination of a target word and a specific L1, since different L1s can have different distances from English for any given word. Darker shading indicates an overlap in points.

Table 10 contains the results of the mixed-models for the Swadesh lists. Surprisingly, there is essentially no effect of distance or of its interaction with L2 proficiency, as the associated effect sizes are almost exactly zero (B = -0.01–0.00, corresponding to IRR = 0.99–1.00). Given this, and given that the associated SEs are also very small (≤0.01 for both B and IRR), this lack of effect is robust within this sample. In addition, there is almost no variance between the L1s based on the associated random effect (SD ≤ 0.03), which suggests that speakers of different L1 used the target words in similar rates.

By contrast, the random effects of *task* and *word* are stronger than the *L1* effect by an order of magnitude or more (SD = 0.33–0.46), and the *task:word* effect is even stronger (SD = 1.36–1.84), which shows that these factors, and primarily the need to use specific words in specific tasks, have a much stronger influence on learners' rate of use of L2 words. Similarly, *frequency* as a control variable also has a very strong effect (B = 3.16–3.30, corresponding to IRR = 23.53–26.99), which was expected since the response variable is a type of frequency measure.

Table 10. Results of the mixed-models, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-10.32	0.16	0.00	<0.01	-65.40	<.001	-9.86	0.14	0.00	<0.01	-68.45	<.001
Distance	-0.01	0.01	0.99	0.01	-1.17	.243	-0.01	0.01	0.99	0.01	-0.36	.718
Proficiency	-0.04	0.02	0.96	0.02	-2.12	.034	0.00	0.02	1.00	0.02	-0.22	.829
Frequency	3.30	0.21	26.99	5.66	15.70	<.001	3.16	0.19	23.53	4.50	16.50	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	0.61	.543	0.00	<0.01	1.00	<0.01	-1.28	.202
Random effects												
Learner_τ_{00}	0.07						0.23					
Task_τ_{00}	0.40						0.33					
Word_τ_{00}	0.38						0.46					
Task:Word_τ_{00}	1.84						1.36					
L1_τ_{00}	0.02						0.03					
L1.Distance_τ_{11}	0.01						0.03					
L1_ρ_{01}	0.55						-0.14					

Table 11 contains the results of the mixed-models based on the parallel dictionaries. The findings of these models support those of the Swadesh-based models. Specifically, there is essentially no effect of distance or of its interaction with proficiency (B = 0.00–0.01, corresponding to IRR = 1.00–1.01), and the associated SEs are also very small (≤0.01 for both B and IRR). In addition, as in the Swadesh-based models, there is almost no variance based on the *L1* random effect (SD ≤ 0.01).

A minor difference is that there is lower variance in the *task* random effect here (SD = 0.03–0.11), but there is also greater variance based on the *word* and *task:word* effects (SD = 0.45–0.65 and SD = 1.50–2.30 respectively), which supports the overall findings in this regard from the Swadesh models, which is that the need to use specific L2 words in specific tasks strongly influences learners' tendency to use those words. Finally, and as expected, frequency is a substantial predictor here too (B = 2.89–2.97, IRR = 18.08–19.50).

Table 11. Results of the mixed-models, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*
(Intercept)	-12.85	0.06	0.00	<0.01	-207.79	<.001	-12.59	0.05	0.00	<0.01	-243.41	<.001
Distance	0.01	<0.01	1.01	<0.01	1.91	.056	0.01	0.01	1.01	0.01	1.04	.301
Proficiency	0.11	0.01	1.12	0.01	9.22	<.001	0.04	0.01	1.04	0.01	4.29	<.001
Frequency	2.89	0.06	18.08	1.05	49.86	<.001	2.97	0.05	19.50	0.99	58.52	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	1.25	.211	0.00	<0.01	1.00	<0.01	1.09	.276
Random effects												
Learner_τ_{00}	0.03						0.04					
Task_τ_{00}	0.03						0.11					
Word_τ_{00}	0.45						0.65					
Task:Word_τ_{00}	2.30						1.50					
L1_τ_{00}	0.00						0.01					
L1.Distance_τ_{11}	0.01						0.01					
L1_ρ_{01}	0.25						0.81					

The results of the models are summarized in Figure 11, which contains the fixed effects from each model, and which illustrates the lack of effect of lexical distance and of its interaction with L2 proficiency. Furthermore, these results are supported by the comparisons with the baseline models (with no lexical distance), which appear in the supplementary information.

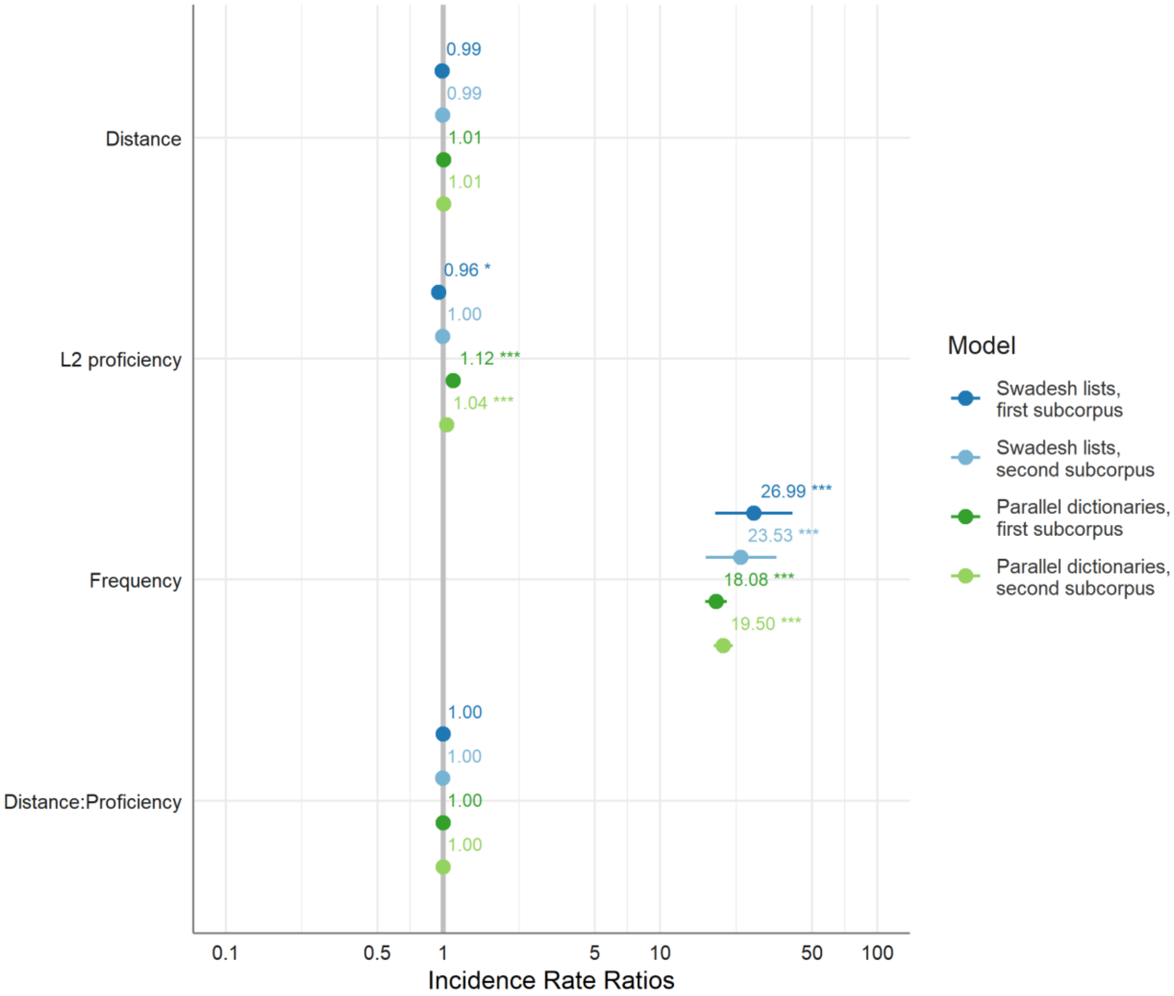


Figure 11. Summary of the models' fixed effects, illustrating the lack of effect of lexical distance and of its interaction with L2 proficiency. *Distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Dots denote the coefficient estimate (in IRR). Lines denote the 95% confidence intervals; where they appear to be missing, it is because they are too narrow to be visible, given the extremely small SEs. Asterisks denote statistical significance of the coefficient estimate (* denotes p < .05 and *** denotes p < .001).

4.4 Discussion

4.4.1 Study summary

We investigated how formal crosslinguistic lexical similarity influences learners' use of L2 vocabulary. Specifically, we investigated whether phonological overlap between L1 words and their L2 counterparts leads to increased use of the L2 words in a task-based setting, and if there is such an effect, whether it is moderated by learners' L2 proficiency. Based on prior research, we expected that learners will be more likely to use L2 words that are similar in form to their L1 counterparts, especially at early L2 proficiency levels.

We found no effect of crosslinguistic similarity on L2 vocabulary use, and no interaction between lexical similarity and L2 proficiency. This null finding was robust across all the combinations of the two subcorpora and two lexical-distance datasets that we examined, since all the associated predictors were tightly clustered around an IRR of 1 (corresponding to a coefficient estimate of 0). In addition, there was very low variance between the L1s based on the associated random effect, which suggests that speakers of different L1 used the target words in similar rates, despite the variation in the average lexical distance between them.[26] Conversely, the *task*, *word*, and especially the *task:word* random effects strongly influenced learner's word choices, which shows that these factors, and primarily the need to use specific words in specific tasks, have a much stronger influence on people's L2 vocabulary choices.

4.4.2 Implications

The main implication of our findings is that the facilitative effect of formal crosslinguistic lexical similarity (in this case, phonological overlap), which relates to cognancy, does not extend to learners' L2 productions to the type of constrained task-based educational setting we examined, which many L2 learners are likely to encounter. This is regardless of learners' L2 proficiency, and applies to learners at the A1–B2 CEFR range of L2 proficiency, though the complete lack of interaction between lexical similarity and L2 proficiency that we found suggests that this likely applies also to learners at the C1–C2 range of proficiency.

This finding supports the finding of Crossley and McNamara (2011) regarding lexical intergroup homogeneity among speakers of different L1s in task-based settings. This suggests

[26] However, the magnitude of this effect should be viewed with caution, since it is likely somewhat underestimated here, due to the small number of L1s involved, particularly in the parallel-based models. Nevertheless, the Swadesh-based models had similar results in this regard, and the exact magnitude of this effect is not crucial to our study, since our focus is on the effects of lexical distance, and as shown in the supplementary information, the estimates for the other effects remain functionally identical when the L1 random effect is removed.

that the lack of L1 effect that they found is not due to their use of a global lexical measure (lexical diversity) and is not due to an idiosyncrasy in their sample, but is rather more likely to be a general feature of task-based settings.

At the same time, however, this does not necessarily contradict studies that found an L1 effect on L2 word choice independently of crosslinguistic lexical similarity (e.g., in stylometry). Rather, the difference may be because the L1 effect that they identified was driven by factors other than crosslinguistic similarity, such as a strong cultural preference for certain words, or because there were weaker task effects in their samples (which also included spontaneous productions), something that also applies to the findings of Rabinovich et al. (2018).

Our finding also does not necessarily contradict the studies that found an effect of lexical similarity on the processing of individual L2 words or on broad L2 acquisition. Rather, it shows that this effect is constrained when it comes to L2 production (i.e., word choice), which is primarily driven by the message the learner wishes to communicate. This is supported by the strong effects of *task*, *word*, and *task*:*word* on word choice, which suggest that the need to use a specific word for a specific task is what drives learners' decision of whether to use it, regardless of whether the word is similar to their L1.

Accordingly, although L2 words that are similar to their L1 translation are likely easier for learners to process and use, the communicative needs of tasks can override this crosslinguistic influence, and drive learners to use necessary words rather than easier ones. Given that, in the present sample, the majority of L2 words appear to be substantially dissimilar from their L1 translations (even for relatively similar L1s, such as German), it seems likely that learners adapt to this quickly, as they have to get used to using L2 words even if they are dissimilar from their L1 translations.

In addition, it is likely that other aspects of the tasks and their educational context played a role in overriding transfer in the present sample, and can also play a role in similar tasks and contexts (especially—but not only—educational ones). For example, it is likely that the lessons associated with tasks involved words (i.e., *content*) that learners then used for practice, or that some task prompts elicited the use of a specific register (i.e., *style*) that necessitated the use of certain words. This supports and extends limited past research which found that factors such as formality and task type may prompt, rather than override, transfer (Jarvis & Pavlenko, 2008), and highlights the potential influence of these situational and contextual factors. This is important, since most transfer research focused on factors pertaining to the associated languages, such as the typological relation between them, or to the learners themselves, such as their L2 proficiency (Jarvis & Pavlenko, 2008).

Finally, note that past studies on the EFCAMDAT found L1 effects in a wide range of linguistic phenomena, including clause subordination (Chen et al., 2021), relative clauses (Alexopoulou et al., 2015), clause-initial prepositional phrases (X. Jiang et al., 2014), grammatical morphemes (Murakami, 2016), articles (Shatz, 2017), and capitalization (Shatz, 2019). Furthermore, X. Jiang et al., (2014) even found evidence of other types of lexical transfer than the one examined here, on the usage rates of certain punctuation marks (e.g., dashes) and phrases (e.g., "to my mind"), though they provide limited information on this. Potential explanations for why these types of transfer occur, while the present one does not, include the following:

- It might be that these types of transfer do not interfere with the communicative needs and task effects in the first place, for example because they have to do primarily with functional elements rather than content words.
- It might be that these types of transfer are "stronger" from a cognitive perspective, and therefore more difficult for communicative needs and task effects to override.
- It might be that these types of transfer are different in some other way that prevents communicative needs and task effects from overriding them (e.g., because they represent negative—rather than positive—transfer).

These explanations raise various interesting theoretical questions, which can be addressed by future research.

One such question is whether and how negative and positive transfer can be influenced differently by communicative needs. For example, if negative transfer prompts a learner to use an L2 word incorrectly (e.g., with spelling errors), then communicative needs that encourage the correct use of the word might override that transfer, by leading the learner to allocate more cognitive resources to ensuring that they use the word correctly. Conversely, if positive transfer helps a learner use an L2 word correctly, then communicative needs that encourage the correct use of the word should not influence the transfer.

Another relevant theoretical question is whether this communicative/task override plays a bigger role when it comes to certain linguistic domains (e.g., lexical vs. syntactic). We hypothesize that this is likely the case, given the contrast between the present null findings and the past positive findings regarding other L1 effects, and given that some aspects of language play a bigger role than others when it comes to achieving communicative/task goals. For example, if a speaker wants to convey the meaning “I ate an apple”, being able to use the word “apple” is generally more important than being able to use the article “an”, since “I ate apple” conveys the original meaning more clearly than “I ate an”.

4.4.3 Limitations and future research

This study’s key limitation is that it relies on a single (albeit large-scale) learner sample. As such, the analyses should be replicated on other samples, in order to determine the generalizability of the findings. Such replications could, for example, analyze the writing of similar learners, of higher-proficiency learners (CEFR C1–C2), or of learners of an L2 other than English (which is a lingua franca). In particular, it would also be good to see analyses on a range of more open-ended tasks, where communicative needs and task effects are not expected to influence word choice as much as in the present study.

In addition, to confirm the findings and determine their generalizability, it will also be beneficial to replicate them using other types of lexical distances or using other lexical-distance datasets. An interesting direction for future research in this regard is to focus on the preference for cognates within synsets, in a manner similar to Rabinovich et al. (2018). This can be done by comparing, within each synset, the probability that speakers of different L1s will use any given word in the synset, and then seeing if their choices reflect a preference for cognates (see the supplementary information for some more details on this).

Finally, future research could address the theoretical questions outlined in §4.4.2, for example by comparing the effects of communicative needs on lexical transfer to their effects on other types of transfer (e.g., syntactic), to further our understanding of how such needs can influence and override transfer.

4.4.4 Conclusions

Our findings suggest that formal lexical similarity—which relates to cognancy and which is measured here based on phonological overlap between words in learners’ L1 and their L2 translations—does not influence learners’ L2 word choice, regardless of learners’ L2 proficiency, in the present task-based educational settings. This suggests that the facilitative effects of formal lexical similarity—the most notable of which is the cognate facilitation effect—are more constrained than expected, and that communicative needs and task effects can override the influence of positive lexical transfer in some cases. Furthermore, this raises questions regarding when and how communicative needs and task effects influence language transfer, for example when it comes to different types of transfer (e.g., positive vs. negative, or lexical vs. syntactic).

5 GENERAL DISCUSSION

5.1 Summary of key research and findings

This research examines the influence of crosslinguistic similarity on lexical transfer from learners' L1 to their target L2. It does so using a broad sample, which covers a large number of learners and texts, and a wide range of words, tasks, L2 proficiency levels, and typologically diverse L1s. In addition, it uses comprehensive mixed-effects models to make full use of the broad sample, and to control for many relevant variables, such as task effects.

The research involves two key studies, which address this topic in a complementary manner. Specifically, the first study investigates whether overall lexical similarity between learners' L1 and their target L2 influences their L2 lexical diversity, while the second study investigates whether lexical similarity between corresponding L1-L2 words influences the usage of the L2 words. As such, the first study examines global measures of lexical similarity and L2 use, and the second study "zooms in" on this potential association and uses a more fine-grained and local lexical measure. Furthermore, both studies examine whether this potential crosslinguistic influence is moderated by learners' L2 proficiency, with the expectation that the influence will be strongest at lower L2 proficiency levels.

Surprisingly, the studies found no effect of lexical similarity on either L2 lexical diversity or word choice, regardless of learners' L2 proficiency. Conversely, there were strong task effects when it comes to both lexical measures.

5.2 Key theoretical implications and future directions

The key finding of this research—the lack of an L1 effect on L2 lexical diversity and word choice—does *not* necessarily contradict past findings on the topic, but rather extends them. It does so by showing that the facilitative effect of lexical similarity (which relates to cognancy) that was found in past research is constrained, and does *not* extend to learners' L2 productions in certain task-based settings, which they are likely to encounter in places such as school. Essentially, it is likely that learners still find it easier to process and use the L2 words that are similar to (and potentially cognate with) their L1, but this does not influence learners' decision of which words to use in the present task-based setting, where learners' lexical choices are driven primarily by constrained communicative needs and associated task effects. This suggests that communicative needs and other task effects can override positive lexical transfer from learners' L1, for example when the need to use a certain word to achieve a certain communicative goal pushes them to use that specific word, rather than an alternative that is slightly easier for them to recall.

As discussed in detail in the two studies, and particularly in study 2, the null findings here stand in stark contrast to findings on other L1 effects in the EFCAMDAT (Alexopoulou et al., 2015; Chen et al., 2021; X. Jiang et al., 2014; Murakami, 2016; Shatz, 2017, 2019). This raises many questions regarding the ability of communicative needs and other task effects to override language transfer, which future research can address.

One key question in this regard is how communicative needs and task effects can influence and override transfer in different linguistic domains (e.g., lexical vs. syntactic). Specifically, while the present findings show how these situational factors can override—and consequently hide—positive lexical transfer, these factors may also play a role when it comes to other types of transfer. For example, these factors may also override potential negative transfer in some cases, in situations where communicative needs prompt speakers to allocate extra cognitive resources to ensuring that a certain linguistic structure is used correctly (e.g., by inhibiting the L1 and relying primarily on the L2 instead). Furthermore, these factors may also influence transfer in other ways, such as by weakening it, rather than overriding it entirely, or by amplifying it, for example by encouraging people to allocate more cognitive resources to a certain structure, in a way that helps them notice additional crosslinguistic similarities, and consequently benefit more from positive transfer than they otherwise would.

Given the null findings in the present lexical domain within the EFCAMDAT sample, and the significant findings in various other linguistic domains, it seems likely that communicative needs and task effects play a bigger role when it comes to transfer that has to do with key communicative needs. For example, as noted in study 2, if a speaker wants to convey the meaning “I ate an apple”, being able to use the word “apple” is generally more important than being able to use the article “an”, since “I ate ~~an~~ apple” conveys the original meaning of the sentence more clearly than “I ate an ~~apple~~”. Similarly, there are many other types of syntactic, semantic, morphological, phonological, and orthographic errors that can be made in this case without substantially interfering with the key meaning of the sentence, such as adding an unnecessary plurality marker to its object (“I ate an apples”) or misspelling its object (e.g., “I ate an applle”). However, the key word in question (“apple”) must be used in a recognizable manner in order to convey the key meaning of the sentence.

Nevertheless, communicative needs can also have to do with the *form* of the language, rather than its *content*. For example, this can be the case when it comes to the *formality* of utterances, which can influence people's word choice in a way that overrides both negative and positive transfer in high-stakes situations, where displaying the right degree of formality is crucial to avoiding serious social conflict.

The ability of communicative needs and task effects to influence and override different types of transfer can be most aptly studied in the future by directly comparing its influence on different types of transfer within the same language sample, or by comparing the existence and magnitude of similar types of transfer across different contexts of language use. For example, this can include running an experiment that involves different levels of task-based constraints, from narrow prompts to more open-ended ones, and investigating to what degree word choice is influenced by crosslinguistic similarity in the different contexts. This direction of research is also important when it comes to investigating the possibility that past studies on transfer may have missed existing transfer or mis-estimated its magnitude, due to the influence of these situational factors.

Finally, from a broader perspective, these findings also emphasize the role of this type of situational factors when it comes to transfer.[27] This is important, since most past transfer research focused on other types of factors, and primarily those pertaining to the linguistic structures in question, for example in terms of their instantiation and frequency across languages, as well as on factors pertaining to the learners themselves, especially when it comes to their language proficiency (James, 2012; Jarvis, 2009; Jarvis & Pavlenko, 2008; Yi, 2012).

5.3 Additional insights

This research also includes additional important contributions beyond the aforementioned key findings.

First, this includes insights into the development, sharing, and use of large-scale language datasets, based primarily on the *EFCAMDAT Cleaned Subcorpus*, which I developed, shared, and used here. This also includes insights into using the ASJP and IDS to calculate lexical distance, and the resulting datasets were also made openly available for use by others,

[27] I use the term "situational factors" rather broadly here, to refer to things such as task effects, but distinctions can be drawn between different types of such factors. For example, Jarvis and Pavlenko (2008) draw a distinction between *social* factors (e.g., idiolect), *situational* factors (e.g., formality), *contextual* factors (e.g., interlocutor), and *performance-related* factors (e.g., task type).

together with all the relevant data, code, and explanations of the development process. I hope that sharing all this will support and encourage the development, sharing, and use of similar language datasets by others. This is especially important given that the opportunities to develop such datasets are increasing rapidly, due to the growing use of digital platforms for various purposes, which generates large amounts of raw data, and given the rapid development of accessible computational tools that can be used to work with such data in an effective and scalable way.

In addition, all the other data and code from the studies is also shared publicly together with the study, in support of open research practices. I am doing this to facilitate the interpretation and replication of the present research, and to raise awareness of data sharing and encourage others to do the same. I believe that this kind of open research is crucial for linguistics and for science in general, and should become the norm and the default option wherever possible.

Finally, this research contains other contributions with significant implications pertaining to language learning, teaching, assessment, and research. This includes:

- Extensive information on L2 lexical diversity, including its developmental patterns across L2 proficiency, its within-level and between-level variance, and the magnitude of the associated task effects. To my knowledge, my research on this is the largest study on the topic, in terms of the combination of the size and diversity of the sample (i.e., the combination of the number of learners and texts, number of tasks, number and typological diversity of the L1s, and the range of the L2 proficiency levels), as well as the scope of the analyses (particularly the use of the comprehensive mixed-effects models).
- Extensive information on the use of lexical distance—and particularly Levenshtein distance and its normalized form—both in the context of general language research, as well as in the context of SLA research in particular. To my knowledge, the review that I present on the topic here is the most comprehensive to date in the context of SLA research, and can therefore facilitate the understanding and use of this and similar measures by other researchers.
- Extensive information on the use of mixed-effects models for SLA research. Though these models are not new to the field, many researchers do not have sufficient familiarity with them, so they remain underutilized, or are used in inappropriate ways (e.g., without proper assumption checking). This appears to be particularly the case when it comes to mixed-models outside of linear mixed-models (LMMs), as in the case of Poisson models. I hope

that by including such models in my work, and providing many details on their use, I will help increase their adoption and proper use in the field.

5.4 Additional future research

In §5.2, I outline the key directions that future research can follow in order to explore the theoretical implications of the present findings (i.e., the way situational factors, such as communicative needs and task effects, can influence and override different types of transfer). In addition, future research can also replicate and extend the present research, by conducting it with one or more of the following modifications:

- Other learner samples (e.g., the Cambridge Learner Corpus instead of the EFCAMDAT).
- Other L1s or a different L2.
- Other measures of lexical distance (e.g., cognancy judgments) or other lexical distance datasets (e.g., the NorthEuraLex).
- Other measures of L2 lexical outcomes (e.g., lexical sophistication).

Of these, I believe that it is most important to replicate this research on other learner samples, in order to determine the generalizability of the findings in other L2 contexts, and provide further insights into how and when communicative needs and task effects override lexical transfer.[28] A different L2 would also be particularly valuable to examine, given English's status as a global lingua franca, though this status also means that many of the relevant datasets use English as an L2, and that acquisition patterns that pertain to it in generally have greater practical implications than for most other languages.

There is, of course, also value in replicating this research using other L1s, other lexical distance measures, other lexical-distance datasets, and other types of L2 outcomes. However, given the robustness of the results within the present analyses, and given the correlation between many of these different measures and datasets (e.g., between different measures of

[28] Although it is also important to emphasize the many strengths of the learner sample that was used here. These include the scale and broadness of the sample (in terms of number of texts, learners, tasks, L1s, and the L2 proficiency levels), as well as the extensive past research that used this sample to study L1 and task effects (Alexopoulou et al., 2015, 2017; Chen et al., 2021; X. Jiang et al., 2014; Michel et al., 2019; Murakami, 2014, 2016; Shatz, 2017, 2019). Given this, and given the robustness of the findings within the present learner sample, *and* given that L1 effects in this sample have been shown to strongly correlate with those in other learner samples (most notably, the Cambridge Learner Corpus, as shown in Murakami, 2013), it is likely that these findings will generalize to other samples that share the same structure, in terms of communicative needs and task effects. What would be most interesting, therefore, is to see how these findings generalize to learner samples with distinctly different structures, such as those that use much more open-ended tasks.

lexical distance), it appears less likely that this would make as much of a difference as replicating these analyses on different learner samples.

Nevertheless, I think there are several interesting and valuable directions for future research in this regard, as well as when it comes to several other associated directions. Below, I briefly outline some of these directions, and present further relevant information, based on various things that I encountered and considered during my PhD research, in the hope that it will inform future work.

5.4.1 Using other lexical-distance datasets

In terms of lexical-distance datasets to use in future research, the *NorthEuraLex* appears to be a highly promising source, in terms of the number and diversity of languages that it covers, the number of words that it contains per language, and the fact that it contains both orthographic and phonetic transcriptions for each word (Dellert et al., 2020).[29] This is a relatively new resource, which was only recently published, and which, at the time of writing this, is awaiting a major update that will involve many improvements and corrections (going from version 0.9 to 1.0), which is why I did not use it in the present research.[30] I believe that once the new major version of this dataset is released, it will be the best lexical-distance dataset to use in follow-up analyses on the present research. I also think that there will be value in establishing a script for automatically generating lexical-distance data from the NorthEuraLex, particularly if it becomes the dataset of choice for most research involving this type of lexical distance.

In addition, another newly released dataset that seems promising is *CogNet* (Batsuren et al., 2021). It has an entirely different structure, since it focuses on assessing cognancy across languages. This can be useful for those interested in using cognancy as a measure of lexical distance, and once the data regarding cognates is extracted, it should also be possible to calculate crosslinguistic lexical distances between the cognates using similar approaches as in the present study.

[29] It also appears to have fewer issues with its transcriptions than the IDS, though this is based only on my initial impression, rather than on comprehensive work with the dataset.

[30] Specifically, the following notice is listed on the project’s homepage: “The current versions of the wordlists have been compiled by non-experts based on available resources, and are therefore guaranteed to contain many errors and inaccuracies. Therefore, they are not adequate for use as a primary reference or data source for any of the languages concerned, but only in computational frameworks where some noise can be dealt with. The next major version (planned for autumn 2020) will contain at least 80 additional languages, a first batch of etymological annotations for the larger families, as well as many updates and corrections based on the feedback of experts and native speakers.” (http://web.archive.org/web/20210828091842/http://northeuralex.org/ as archived on 28-Aug-21).

5.4.2 Accounting for orthography

The present research focuses on crosslinguistic *phonological* similarity, and does not account directly for *orthographic* similarity, though the mixed-effects models that were used in the main analyses do account for it indirectly through the *L1* and *word* random effects. Reasons for this include the lack of orthographic data in the ASJP, the problem with calculating orthographic distance between languages with different scripts,[31] and the strong correlation between phonological and orthographic distance.[32] Nevertheless, it could be interesting to expand analyses to account for both phonological and orthographic similarity.

The main way to do this is to include in the sample only L1s that share the same script as the target L2, calculate both orthographic and phonological distances between the word pairs in the sample, and then include both distance measures in the analyses, while ensuring that potential issues (especially collinearity) are addressed.

In addition, another way in which orthographic similarity can be taken into account is through similarity at the *script* level (or *writing system*), rather than at the *word* level. Specifically, I propose that differences in script between learners' L1 and the target L2 can be calculated using three types of scales:

i. **Binary scale**. This involves distinguishing between L1s that use the same script as the target L2 and those that do not.
ii. **Categorical scale**. This involves categorizing the L1s and the L2 based on their script (e.g., Latin, Cyrillic, or Arabic).

[31] The distances are almost always maximal when the scripts are different, and therefore they are largely meaningless. However, note that L1-L2 similarities and differences can have *cross-script* effects (i.e., they can influence learners' L2 even when the two languages use different script), which means, for example, that the facilitative effect of cognate status on L2 word use can occur even when the L1 and L2 use different scripts (Bowers et al., 2000; Bowers & Michita, 1998; Gollan et al., 1997; Hoshino & Kroll, 2008; N. Jiang, 1999; Kroll et al., 2012; Muljani et al., 1998; Thierry & Wu, 2007; Wu & Thierry, 2010), though differences in writing system do play a role in some cases (J. Zhang et al., 2019).

[32] Specifically, in the parallel dictionaries, where all the L1s share English's Latin script, and where there are orthographic transcriptions, there was a strong correlation between phonological and orthographic distance, both in the case of LDN (r = .68, *95% CI* = [.67, .70], p < .001), and in the case of LD (r = .73, *95% CI* = [.71, .74], p < .001). Similar strong correlations have been found in other studies, such as in Carrasco-Ortiz et al. (2021), who found a correlation of r = .782 between phonological and orthographic distance in a dataset of English and Spanish words. Furthermore, this correlation has been raised as a problematic source of collinearity, leading some researchers to omit orthographic distance from their statistical analyses, and keep only phonological distance (e.g., De Wilde et al., 2020, 2021). This does not entirely rule out the inclusion of both distances types, as the issue of collinearity can potentially be mitigated through the use of sufficiently large and diverse samples, but it does mean that researchers who consider including both distance measures in their analyses should do so with caution (Morrissey & Ruxton, 2018; R. M. O'Brien, 2007; Winter, 2019).

iii. **Ordinal scale**. This involves categorizing the L1s based on their script, and then ranking them based on how similar their script is to that of the L2, based on the type of script used (e.g., phoneme-based vs. syllable-based) and potentially also the degree of overlap in symbols between the scripts.

In the case of the ordinal scale, there it is possible to use the following four categories, described here in the context of English as the target L2:

i. **Same script.** This includes languages such as French and Spanish, which use a script that is identical or almost identical to English.

ii. **Different script with overlap.** This includes languages such as Russian and Greek, which use a substantially different script than English, but which contain some overlap in terms of certain graphemes (e.g., /a/ in Russian).[33]

iii. **Different script without overlap.** This includes languages such as Hebrew and Arabic, which use a script that does not overlap with English at all in terms of graphemes, but which is the same type of script overall (i.e., an alphabet).

iv. **Different script type.** This includes languages such as Chinese and Japanese, which contain no overlap in graphemes with English, and which use a non-alphabetic script.[34] Note that, in an *alphabetic system*, each grapheme generally corresponds to a phoneme, so non-alphabetic scripts are generally either a *syllabary*, if graphemes correspond to syllables, or a *logographic system*, if graphemes correspond to morphemes.

This type of crosslinguistic similarity be taken into account both in analyses that look at similarities at the word level, as well as those that look at similarities at the language level.[35]

[33] It is possible to refine this category of the scale further, for example by calculating the proportion of the alphabet which overlaps. In this regard, it may also be possible to refine the previous category of the scale in a similar manner.

[34] Though speakers of these languages may still have substantial experience with the Latin script, for example through the Rōmaji script in Japanese, which can complicate analyses that use this scale.

[35] In the present research, the effect of script is broadly taken into account through the *L1* random effect, but is largely irrelevant for most of the analyses. Specifically, in the case of study 1 (on lexical diversity), similarity in script is almost perfectly correlated with phonological distance, while in the case of study 2 (on word choice), all the L1s that are included in the parallel dictionaries share the same script, so it could only make a difference for the analyses on the Swadesh lists there.

5.4.3 Using phonological weights

This refers to assigning different weights to the transformation of different phonological units. For extensive information on this, see the section on “Feature edit distance” in the supplementary information of the word-choice study (Appendix E).

5.4.4 Accounting for diacritics

Diacritics are marks that modify base characters in orthographic or phonological transcriptions, as in the case of the /ˆ/ in /ê/ (Ball, 2001). Diacritics are similar to other types of symbols that are used to modify base characters, such as *suprasegmentals* (e.g., the *long* marker /ː/), which here will all be grouped under the term “diacritics”, due to the similarities in how they are represented and how they function.

Diacritics pose two main challenges to calculations of lexical distance.[36] First, differences in diacritics may be less substantial than differences in base characters, for example in terms of how they are perceived by language speakers (Heeringa et al., 2013). Second, there is often inconsistency in how diacritics are transcribed, both between and within transcribers/datasets (McSweeny & Shriberg, 1995; Ramsdell et al., 2007; Shriberg & Lof., 1991). Nevertheless, I included diacritics in my calculations of lexical distance without treating them differently than other characters, for several reasons.

First, most past research that used and validated LD(N) did not discuss the handling of diacritics (e.g., Petroni & Serva, 2008; Schepens et al., 2012; Wichmann, 2019), which suggests that this research did not treat diacritics differently than other characters, so this course of options seems a reasonable and conservative option to use here. This choice is supported by the fact that the language-level distances that we found in the analyses here align with the distances that are expected based on general language classification, as shown in the two main studies of the thesis.

Another reason for treating diacritics the same as other characters when calculating lexical distance is that there is currently no way to handle diacritics that is clearly better than this. Specifically:

- One option is to weigh diacritic-based changes differently than character-based changes, for example by assigning a weight of 0.5 to a transformation of /e/ into /é/,

[36] This also applies to calculations of phonological or orthographic overlap or similarity, which often serve as a proxy for lexical distance, as in the present research.

compared to a weight of 1 if it is transformed into /a/ (Heeringa et al., 2013; Silveira & Leussen, 2015). However, this does not fully address the issue of inconsistency in the transcription of diacritics, and it is not clear that diacritic-based changes are indeed less substantial than character-based changes in all case (e.g., when assessing psychotypological distance). Furthermore, this adds substantial complexity and arbitrariness to the analyses, for example when deciding how diacritics should be weighted compared to other character transformations (e.g., 0.5? 0.3? 0.25?), which can lead to various issues with running such analyses in a replicable manner.

- Another option is to remove or ignore diacritics entirely when calculating lexical distance, so that only the base characters in each string are compared (e.g., Jäger, 2018; Luján-Mora & Palomar, 2001; Sanders & Chin, 2009; Santos et al., 2018; Wieling et al., 2007, 2014; Zhang, 2018). This has the benefit of solving most issues associated with diacritics, but it comes directly at the cost of lost linguistic information. Furthermore, it also adds substantial complexity to the analyses, especially when different types of diacritics need to be removed, since this is not always a straightforward process, as we will see soon.
- Another option is to use FD(N) instead of LD(N), since, as outlined in the previous sub-section, such distances may be better able to handle diacritics, by properly parsing the linguistic information associated with them. However, in addition to the general issues associated with FD(N) that were discussed previously, this still does not fully address the issue of inconsistency in diacritic transcriptions, and this can also lead to other issues, such as inconsistency in terms of how FD(N) calculations treat different types of diacritics.[37] In addition, FD(N), at least in its current form, generally only applies to phonological transcriptions, but diacritics may also appear in orthographic transcriptions.

The complexity of trying to deal with diacritics is compounded when considering the many different types of diacritics, which have different linguistic functions and take different programmatic forms. For example, focusing on the programmatic form of diacritics, you can have:

[37] For example, PanPhon acknowledges the diacritic in /ɔ̃/, but ignores the diacritic in /ž/.

- Symbol-based diacritics that must be attached to a base character, meaning that deleting the base character also generally deletes the diacritics in software (e.g., the nasalization diacritic in /ɔ̃/).
- Symbol-based diacritics that must be attached to a base character, and that modify multiple (usually two) characters (e.g., the affricate mark in /t͡s/).
- Symbol-based diacritics that can appear on their own (e.g., the long diacritic in /aː/).
- Letter-based diacritics that can appear on their own (e.g., the palatalization diacritic in /pʲ/).
- *Precomposed characters* (also known as *precombined glyphs*), which are a combination of a diacritic and base character(s) that is transcribed as a single character (e.g., /ą/, which replaces the 2-character /a̧/, or /ʧ/, which replaces the 3-character /t͡ʃ/).

This variability is important, since solutions for handling diacritics may work with some of them, but not others, which can increase the room for error in any analyses involved.[38]

Overall, diacritics may lead to issues in calculations of lexical distance, since diacritic-based differences *may* be perceived as less substantial than character-based differences, and since the transcription of diacritics is often more inconsistent than the transcription of base characters. Nevertheless, these issues do *not* appear to invalidate the use of LD(N) when calculating lexical distance, especially when it comes to large-scale analyses that are expected to accommodate this type of limited noise. Furthermore, there is currently no approach that is clearly superior to simply keeping the diacritics in the dataset and treating them the same way as other characters, since all the alternatives also involve various issues.

As such, my goal in writing about this here is to raise awareness of the influence of diacritics on lexical distance, and to provide an initial overview of this concept, the issues involved, and the current solutions. This will help researchers who work with lexical distance to at least be aware of the issue, so they can make an informed decision regarding how to handle

[38] For example, when it comes to Python, solutions that are often recommended for removing diacritics are the *unidecode* function in the *unidecode* library (Solc, 2019) and the *normalize* function in the *unicodedata* module (*What Is the Best Way to Remove Accents in a Python Unicode String?*, 2009). However, these methods fail to account for non-accent diacritics, such as the labialization symbol (/ʷ/), which they change into a /w/, or the long symbol (/ː/), which they simply do not remove. Furthermore, these solutions can also involve other issues, such as *lossy transliteration* of Unicode characters into ASCII, which means that characters that exist in the Unicode character-set but not in the more limited ASCII character-set are not preserved during the transliteration process. This means, for example, that /ɔ̃/ might be converted to /o/, as in the case of the *unidecode* function, or deleted entirely, as in the case of the relevant function in the *unicodedata* library. These issues will not apply to every language sample, and it is possible to address them by identifying and removing problem characters using *regex*, while potentially combining this with more appropriate functions, such as the *deaccent* function from the *gensim* library in Python (Rehurek & Sojka, 2010). Nevertheless, this further illustrates some of the pitfalls and complexity associated with handling diacritics.

diacritics in their analyses. In addition, this can prompt further discussions on the topic, which may lead to the development of better approaches for handling diacritics in linguistic analyses.

5.4.5 Assessing segmental frequency and permissibility

The present research, as well as most research on the influence of crosslinguistic similarity on lexical transfer, focuses on the similarity between corresponding individual words, including when the mean distance between words in different languages is used. An interesting direction for future research is to examine the influence of phonological and orthographic similarity between languages on lexical transfer, in terms of permissibility and frequency of phonological and orthographic segmental units in each language (e.g., a phoneme such as /b/, or a combination of phonemes such as /ba/). This type of system similarity has been conceptualized in various ways, such as through *phonotactic typicality*, which expresses the degree to which the phonological structure of L2 words resembles the phonological structure of a learner's L1 (de Groot, 2006), and this type of crosslinguistic similarities and differences can influence L2 processing, acquisition, and use (N. C. Ellis & Beaton, 1993; French & O'Brien, 2008; Llach, 2010; Llach et al., 2006; Martin & Ellis, 2012; I. O'Brien, 1998; I. O'Brien et al., 2006; Speciale et al., 2004).

For example, when L2 words contain sounds that are impermissible in learners' L1, learners often modify those sounds in various ways instead of producing them faithfully (Carlisle, 1991, 1997; Carlson, 2018a; Carlson et al., 2016; Davidson, 2011; Dupoux et al., 2011). This means, for instance, that since word-initial /s/ clusters are impermissible in Spanish, when Spanish speakers encounter such clusters in languages such as English, they tend to insert a word-initial /e/ in order to repair the seemingly illicit cluster (e.g. *spirit* → *espirit*), especially during the initial stages of acquisition (Carlisle, 1991, 1997; Carlson, 2018b). This has been attributed to parallel activation of L2 and L1 phonotactic constraints during the processing of the L2 words (Carlson, 2018a; Carlson et al., 2016; Freeman et al., 2016, 2017). This issue can require learners to dedicate more resources to processing, acquiring, and using L2 words with the impermissible clusters, which can interfere with other various aspects of L2 lexical development, such as L1-L2 mapping and the acquisition of other L2 words (Carlson, 2018a; Carlson et al., 2016; Davidson, 2011; Dupoux et al., 2011).

There are many other cases of this type of crosslinguistic influence. For example, learners are better able to recall phototactically typical words compared to atypical words, which is attributed to the increased L1-L2 similarity in phonological encoding making it easier for learners to generate phonological coding of the L2 word forms in phototactically typical words (de Groot, 2006; de Groot & van Hell, 2005). In addition, the L1-L2 similarity of phonotactic patterns can also affect the *pronounceability* of L2 words, with words that are more similar being viewed as more pronounceable (N. C. Ellis & Beaton, 1993). Finally, orthographic similarity can influence lexical transfer, for example by facilitating L2 lexical processing in cases where learners can establish beneficial associations between the writing systems of the L1 and the L2 (Koda, 1996; Muljani et al., 1998).

Overall, research shows that phonological and orthographic similarity at the language level between learners' L1 and their target L2 can influence L2 processing, acquisition, and use, even outside the context of cognancy or similarity between specific corresponding words. However, as with the cognate facilitation effect, most research on this focused on narrow assessments of L2 processing, so there is a need for broad assessments of L2 production in this regard.

There is a comprehensive recent study on a similar concept by Schepens et al. (2020), which looked at phonological similarities between languages based on their phonological inventories, and examined their influence on a composite score of L2 proficiency. However, what I think would be interesting is to go beyond the basic inventory, and look at all the available combinations of sounds and symbols within the languages, while considering their permissibility and frequency in each language, for example based on their n-grams within the language.[39] This can be based on wordlists such as those in the NorthEuraLex, or on wordlists that also include the frequency of each word in the language, such as those in the *wordfreq* package in Python. Using this data, regarding how often combinations of sounds/symbols appear in learners' L1 and their target L2, it would be interesting to answer the following questions about lexical transfer:

- Does the frequency of L1 phonological/orthographic units influence their frequency in learners' L2 productions? If so, does this occur due to a preference for units that are similar across the languages, avoidance of units that are different, or both?
- Is the influence of frequency distinct from that of permissibility? Specifically, it might be the case that as long as a cluster is permissible in the L1 (i.e., has a frequency greater

[39] This aligns with some other suggestions for ways in which n-grams can be used to quantify distance between languages (Gamallo et al., 2017).

than zero), then its frequency has no effect on L2 productions. Conversely, it might be the case that permissibility does not matter, and all the matters is L1 frequency, where impermissibility is merely a case where frequency is equal to zero. In addition, it might also be the case that both frequency and permissibility matter, (i.e., that frequency influences acquisition, but the effect of the transition from low-frequency to impermissibility is inconsistent with the effect that is expected based on a simple decrease in frequency).

Furthermore, these analyses can be refined by looking not only at general frequency and permissibility, but also at frequency and permissibility given the position of segmental units within words or in relation to other segmental units.[40]

5.4.6 Assessing other L2 outcomes

This research examines the influence of crosslinguistic similarity on lexical transfer in the context of two L2 outcomes: lexical diversity and the usage rates of individual words. Future research can expand on this by looking other types of L2 outcomes.

For example, when it comes to other global lexical measures, it would be interesting to look at other types of measures beyond lexical diversity (Jarvis, 2013; Mazgutova & Kormos, 2015), such as lexical sophistication (Kyle & Crossley, 2016) and lexical proficiency (Baba, 2009). On the other hand, when it comes to the usage patterns of individual words, it would be interesting to look at the *emergence* of L2 words, possibly by using *event history analysis* (also known as *survival analysis*)(Hox et al., 2018; Ota & Green, 2013). Furthermore, it would also be interesting to analyze the transfer-based spelling and pronunciation errors that people make as a result of crosslinguistic similarity and lexical transfer, using the same type of large-scale analyses as in the present study.

[40] Specifically, it is possible to take the *absolute* and *relative* position of the units into account (N. C. Ellis & Beaton, 1993). *Absolute position* refers to the position of a segment/sequence within a word, regardless of the position of other segments. For example, if a language does not allow the /t/ sound to appear in word-final position, that phonotactic constraint refers to its absolute position. *Relative position* refers to the position of a segment/sequence within a word, relatively to the position of other segments. For example, if a language does not allow the /t/ sound to appear immediately after an /r/ sound, that phonotactic constraint refers to its relative position.

5.4.7 Other considerations and open questions

When it comes to the influence of crosslinguistic similarity on lexical transfer, there are many related research directions that are worth pursuing, beyond the ones mentioned so far.

One topic of interest is whether normalized and non-normalized distances play a different role in influencing transfer, and whether this is moderated by word length. For example, do two-character transformations in a four-letter word lead to different lexical transfer than three-character transformations in a six-letter word? What about edge cases such as the pair /aaa/-/bbb/ compared to the pair /aaa/-/bbbbbb/, where the LDN is equal but the LD is markedly different?

A related topic of interest is whether there are bounds to the effects of similarity. For example, is it the case that there must be overlap in at least 40% of phonemes in an L1-L2 word pair for learners to benefit from crosslinguistic similarity? Similarly, is it possible that there is no difference between words that are 60% similar and those that are 90%, because past a certain point additional similarity no longer confers additional benefits?

Finally, another topic of interest is how else can other aspects of word complexity, beyond word length, influence lexical transfer? For example, is there a bigger cognate facilitation effect when it comes to complex L2 words (e.g., words that contain difficult phonological clusters) because those words are harder to learn, so learners are more likely to turn to their L1 for help?

Answering these questions will help us understand how crosslinguistic similarity influences L2 lexical development. Furthermore, this will inform the methodology that is used to research the topic, for example by helping us determine what constitutes cognancy from a psycholinguistic perspective. In addition, this can also be beneficial in applied contexts, for example when it comes to understanding when it is worthwhile to highlight crosslinguistic similarities to facilitate the learning of L2 words.

There are, of course, many, many other topics that can also be considered in future research. These include, for example, the role of systematicity and individual variation in lexical transfer (Murakami, 2016), the role of part of speech (N. C. Ellis & Beaton, 1993), and the role of factors such as orthographic depth (Schepens, Dijkstra, et al., 2013). The questions and directions that I am proposing here are merely the ones that I believe to be the most relevant given the focus of the present research.

5.5 Conclusion

This research investigates the effects of crosslinguistic similarity on lexical transfer. It uses a broad learner-corpus sample, containing L2 English texts written by a diverse range of learners in a task-based setting. Interestingly, lexical similarity between languages did not influence L2 lexical diversity, regardless of learners' L2 proficiency. Similarly, lexical similarity between corresponding L1-L2 words did not influence the use of the L2 words, again regardless of L2 proficiency. Conversely, there were strong task effects on both lexical measures.

These findings show that the facilitative effect of lexical similarity is constrained, and suggest that communicative needs and other task effects can override positive lexical transfer. This raises questions regarding when and how communicative needs and other task effects can override and influence transfer, for example when it comes to different types of transfer (e.g., positive vs. negative, or lexical vs. syntactic), and highlights the potential influence of such situational factors on language transfer. In addition, this research contains many insights into related topics, primarily in the form of additional findings and suggestions for future work, for example when it comes to accounting for task effects in language assessment, using online platforms to develop language corpora, and measuring crosslinguistic distance.

6 Bibliography

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research, 1*(1), 96–129. https://doi.org/10.1075/ijlcr.1.1.04ale

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning, 67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Allen, B., & Becker, M. (2015). Learning alternations from surface forms with sublexical phonology. In *Lingbuzz*. http://ling.auf.net/lingbuzz/002503

Altenberg, B., & Granger, S. (2001). The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. *Applied Linguistics*, *22*(2), 173–194. https://doi.org/10.1093/applin/22.2.173

Ard, J., & Homburg, T. (1983). Verification of language transfer. In S. Gass & L. Selinker (Eds.), *Language Transfer in Language Learning* (pp. 157–176). Newbury House.

ASJP. (2018). https://asjp.clld.org

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer Academic Publishers.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. *The Mental Lexicon, 2*(3), 419–463. https://doi.org/10.1075/ml.2.3.06baa

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, *18*(3), 191–208. https://doi.org/10.1016/j.jslw.2009.05.003

Bakker, D., Brown, C. H., Brown, P., Egorov, D., Grant, A., Holman, E. W., Mailhammer, R., Müller, A., Velupillai, V., & Wichmann, S. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, *13*(1), 169–181. https://doi.org/10.1515/LITY.2009.009

Ball, M. J. (2001). On the status of diacritics. *Journal of the International Phonetic Association, 31*(2), 259–264. https://doi.org/10.1017/S0025100301002067

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639–647.

Barcroft, J. (2004). Second Language Vocabulary Acquisition: A Lexical Input Processing Approach. *Foreign Language Annals*, *37*(2), 200–208. https://doi.org/10.1111/j.1944-9720.2004.tb02193.x

Barrus, T. (2019). *pyspellchecker* (0.5.3). Python library. https://github.com/barrust/pyspellchecker

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batsuren, K., Bella, G., & Giunchiglia, F. (2021). A large and evolving cognate database.

Language Resources and Evaluation. https://doi.org/10.1007/s10579-021-09544-6

Beijering, K., Gooskens, C., & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. In M. van Koppen & B. Botma (Eds.), *Linguistics in the Netherlands* (pp. 13–24). John Benjamins. https://doi.org/10.1075/avt.25.05bei

Benson, C. (2002). Transfer / Cross-linguistic influence. *ELT Journal*, *56*(1), 68–70.

Blom, E., Boerma, T., Bosma, E., Cornips, L., van den Heuij, K., & Timmermeister, M. (2020). Cross-language distance influences receptive vocabulary outcomes of bilingual children. *First Language*, *40*(2), 151–171. https://doi.org/10.1177/0142723719892794

Bolker, B. M. (2020). *Post-model-fitting procedures with glmmTMB models: Diagnostics, inference, and model output* (pp. 1–18). https://cran.r-project.org/web/packages/glmmTMB/vignettes/model_evaluation.pdf

Bosma, E., Blom, E., Hoekstra, E., & Versloot, A. (2019). A longitudinal study on the gradual cognate facilitation effect in bilingual children's Frisian receptive vocabulary. *International Journal of Bilingual Education and Bilingualism*, *22*(4), 371–385. https://doi.org/10.1080/13670050.2016.1254152

Bowers, J. S., & Michita, Y. (1998). An investigation into the structure and acquisition of orthographic knowledge: Evidence from cross-script Kanji-Hiragana priming. *Psychonomic Bulletin and Review*, *5*(2), 259–264. https://doi.org/10.3758/BF03212948

Bowers, J. S., Mimouni, Z., & Arguin, M. (2000). Orthography plays a critical role in cognate priming: Evidence from French/English and Arabic/French cognates. *Memory and Cognition*, *28*(8), 1289–1296. https://doi.org/10.3758/BF03211829

Brenders, P., van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology*, *109*(4), 383–396. https://doi.org/10.1016/j.jecp.2011.03.012

Brooks, M. E., Kristensen, K., Benthem, K. J. Van, Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brooks, M. E., Kristensen, K., Darrigo, M. R., Rubim, P., Uriarte, M., Bruna, E., & Bolker, B. M. (2019). Statistical modeling of patterns in annual reproductive rates. *Ecology*, *100*(7), 1–7. https://doi.org/10.1002/ecy.2706

Brown, C. H., Holman, E. W., & Wichmann, S. (2013). Sound correspondences in the world's languages. *Language*, *89*(1), 4–29.

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals*, *61*(4), 285–308. https://doi.org/10.1524/stuf.2008.0026

Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2020). The role of conflicting representations and uncertainty in internal error detection during L2 learning. *Language Learning*, *70*(S2), 75–103. https://doi.org/10.1111/lang.12401

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (Issue May, pp. 35–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

Carlisle, R. S. (1991). The influence of environment on vowel epenthesis in Spanish/English interphonology. *Applied Linguistics*, *12*(1), 76–95. https://doi.org/10.1093/applin/12.1.76

Carlisle, R. S. (1997). The Modification of Onsets in a Markedness Relationship: Testing the Interlanguage Structural Conformity Hypothesis. *Language Learning*, *47*(2), 327–361. https://doi.org/10.1111/0023-8333.101997010

Carlson, M. T. (2018a). Making Room for Second Language Phonotactics: Effects of L2 Learning and Environment on First Language Speech Perception. *Language and Speech*, *61*(4), 598–614. https://doi.org/10.1177/0023830918767208

Carlson, M. T. (2018b). Now you hear it, now you don't: Malleable illusory vowel effects in Spanish-English bilinguals. *Bilingualism*. https://doi.org/10.1017/S136672891800086X

Carlson, M. T., Goldrick, M., Blasingame, M., & Fink, A. (2016). Navigating conflicting phonotactic constraints in bilingual speech perception. *Bilingualism*, *19*(5), 939–954. https://doi.org/10.1017/S1366728915000334

Carrasco-Ortiz, H., Amengual, M., & Gries, S. T. (2021). Cross-language effects of phonological and orthographic similarity in cognate word recognition. *Linguistic Approaches to Bilingualism*, *11*(3), 389–417. https://doi.org/10.1075/lab.18095.car

Carroll, J. B. (1967). On sampling from a lognormal model of word frequency distribution. In H. Kučera & W. N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406–424). Brown University Press.

Carroll, S. E. (1992). On cognates. *Second Language Research*, *8*(2), 93–119.

Casaponsa, A., Antón, E., Pérez, A., & Duñabeitia, J. A. (2015). Foreign language comprehension achievement: Insights from the cognate facilitation effect. *Frontiers in Psychology*, *6*, 1–12. https://doi.org/10.3389/fpsyg.2015.00588

Cebrian, J. (2000). Transferability and Productivity of L1 Rules in Catalan-English Interlanguage. *Studies in Second Language Acquisition*, *22*(01), 1–26. https://doi.org/10.1017/S0272263100001017

Cenoz, J., Leonet, O., & Gorter, D. (2021). Developing cognate awareness through pedagogical translanguaging. *International Journal of Bilingual Education and Bilingualism*, 1–15. https://doi.org/10.1080/13670050.2021.1961675

Chavula, C., & Suleman, H. (2016). Assessing the Impact of Vocabulary Similarity on Multilingual Information Retrieval for Bantu Languages. *Proceedings of the 8th Annual Meeting of the Forum on Information Retrieval Evaluation*, 16–23. https://doi.org/10.1145/3015157.3015160

Chen, X., Alexopoulou, T., & Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behavior Research Methods*, *53*(2), 803–817. https://doi.org/10.3758/s13428-020-01456-7

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

Cop, U., Dirix, N., Van Assche, E., Drieghe, D., & Duyck, W. (2017). Reading a book in one or two languages? An eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism*, *20*(4), 747–769. https://doi.org/10.1017/S1366728916000213

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, *5*, 161–170.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(5), 1283–1296. https://doi.org/10.1037/0278-7393.26.5.1283

Costa, A., Santesteban, M., & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language*, *94*(1), 94–103. https://doi.org/10.1016/j.bandl.2004.12.002

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, *20*(4), 271–285. https://doi.org/10.1016/j.jslw.2011.05.007

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/applin/amt056

Davidson, L. (2011). Phonetic, Phonemic, and Phonological Factors in Cross-Language Discrimination of Phonotactic Contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 270–282. https://doi.org/10.1037/a0020988

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70.

de Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*(3), 463–506. https://doi.org/10.1111/j.1467-9922.2006.00374.x

de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1–56. https://doi.org/10.1111/0023-8333.00110

de Groot, A. M. B., & van Hell, J. G. (2005). The Learning of Foreign Language Vocabulary. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 9–29). Oxford University Press. https://doi.org/http://dx.doi.org/10.4274/Jcrpe.714

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure: How do word-related variables and proficiency influence receptive vocabulary learning? *Language Learning*, *70*(2), 349–381. https://doi.org/10.1111/lang.12380

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Formal versus informal L2 learning: How do individual differences and word-related variables influence French and English L2 vocabulary learning in Dutch-speaking children? *Studies in Second Language Acquisition*, Advance online publication. https://doi.org/10.1017/S0272263121000097

Dean, C. B., & Lundy, E. R. (2016). *Overdispersion*. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06788.pub2

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J., & Jäger, G. (2020).

NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, *54*(1), 273–301. https://doi.org/10.1007/s10579-019-09480-6

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), 1–12. https://doi.org/10.1371/journal.pone.0121945

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*(3), 284–301. https://doi.org/10.1016/j.jml.2009.12.003

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*, *64*(3), 199–210. https://doi.org/10.1016/j.jml.2010.12.004

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2020). *Ethnologue: Languages of the world. Twenty-third edition.* SIL International. https://www.ethnologue.com

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2021). *Ethnologue: Languages of the world (twenty-fourth edition)*. SIL International. https://www.ethnologue.com

Ecke, P. (2015). Parasitic vocabulary acquisition, cross-linguistic influence, and lexical retrieval in multilinguals. *Bilingualism*, *18*(2), 145–162. https://doi.org/10.1017/S1366728913000722

Eden, S. E. (2018). *Measuring phonological distance between languages*. University College London.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic Determinants of Foreign Language Vocabulary Learning. *Language Learning*, *43*(4), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., & Arshanapalli, B. G. (2014). Model selection criterion: AIC and BIC. In *The basics of financial econometrics* (pp. 399–403). https://doi.org/10.1002/9781118856406.app5

Feinerer, I., & Hornik, K. (2018). *tm: Text Mining Package*. R package. https://cran.r-project.org/package=tm

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech, Language, and Hearing Research*, *58*, 840–852. https://doi.org/10.1044/2015

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., & Aumont, X. (2016). Using phonologically weighted Levenshtein distances for the prediction of microscopic intelligibility. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 650–654. https://doi.org/10.21437/Interspeech.2016-431

Forthmann, B., & Doebler, P. (2021). Reliability of researcher capacity estimates and count data dispersion: A comparison of Poisson, negative binomial, and Conway-Maxwell-Poisson models. *Scientometrics*, *126*(4), 3337–3354. https://doi.org/10.1007/s11192-021-03864-8

Freeman, M. R., Blumenfeld, H. K., & Marian, V. (2016). Phonotactic constraints are activated across languages in bilinguals. *Frontiers in Psychology*, *7*(702), 1–12. https://doi.org/10.3389/fpsyg.2016.00702

Freeman, M. R., Blumenfeld, H. K., & Marian, V. (2017). Cross-linguistic phonotactic competition and cognitive control in bilinguals. *Journal of Cognitive Psychology*, *29*(7), 783–794. https://doi.org/10.1080/20445911.2017.1321553

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*(3), 463–487. https://doi.org/10.1017/S0142716408080211

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). From language identification to language distance. *Physica A: Statistical Mechanics and Its Applications*, *484*, 152–162. https://doi.org/10.1016/j.physa.2017.05.011

Geertzen, J., Alexopoulou, T., Baker, R., Jiang, S., & Korhonen, A. (2013). *The EF-Cambridge open language database (EFCAMDAT) user manual part I: Written production*. https://corpus.mml.cam.ac.uk/

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Millar, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, & A. Tseng (Eds.), *Selected Proceedings of the 2012 Second Language Research Forum* (pp. 240–254). Cascadilla Proceedings Project.

Gerard, L. D., & Scarborough, D. L. (1989). Language-Specific Lexical Access of Homographs by Bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 305–315. https://doi.org/10.1037/0278-7393.15.2.305

Gollan, T. H., Forster, K. I., & Frost, R. (1997). crucial for Translation Priming With Different Scripts : Masked Priming With providing and Noncognates in Bilinguals the effect of. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1122–1139. http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=1997-06939-004&S=L&D=pdh&EbscoContent=dGJyMMvl7ESeqa44v%2BbwOLCmr1Cep7NSsKu4TLCWxWXS&ContentCustomer=dGJyMPGot1CzrrZLuePfgeyx44Dt6fIA

Gooskens, C. (2006). Linguistic and extra-linguistic predictors of inter-Scandinavian intelligibility. In J. van de Weijer & B. Los (Eds.), *Linguistics in the Netherlands* (Vol. 23, pp. 101–113). John Benjamins. https://doi.org/10.1075/avt.23.12goo

Gooskens, C., & Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, *16*(3), 189–207. https://doi.org/10.1017/S0954394504163023

Gordon, B., & Caramazza, A. (1982). Lexical decision for open-and closed-class words: Failure to replicate differential frequency sensitivity. *Brain and Language*, *15*(1), 143–160.

Granger, S., & Wynne, M. (1999). Optimising measures of lexical variation in EFL learner corpora. In J. Kirk (Ed.), *Corpora Galore* (pp. 249–257). Rodopi.

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/10.1029/2001gc000192

Greenhill, S. J. (2011). Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, *37*(4), 689–698. https://doi.org/10.1162/COLI_a_00073

Gries, S. T. (2021). (Generalized linear) mixed-effects modeling: A learner corpus example. *Language Learning*, *71*(3), 757–798. https://doi.org/10.1111/lang.12448

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Gudmestad, A. (2012). Acquiring a Variable Structure: An Interlanguage Analysis of Second Language Mood Use in Spanish. *Language Learning*, *62*(2), 373–402. https://doi.org/10.1111/j.1467-9922.2012.00696.x

Gujord, A.-K. H. (2020). Crosslinguistic influence. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 345–357). Routledge. https://doi.org/10.1002/9780470756492.ch15

Hall, C. J. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *International Review of Applied Linguistics in Language Teaching*, *40*(2), 69–87. https://doi.org/10.1515/iral.2002.008

Hall, C. J., & Ecke, P. (2003). Parasitism as a default mechanism in L3 vocabulary acquisition. In B. H. & U. J. J. Cenoz (Ed.), *The multilingual lexicon* (Dordrecht, pp. 71–85).

Hall, K. C., Allen, B., Fry, M., Johnson, K., Lo, R., Mackie, S., & McAuliffe, M. (2017). *Phonological CorpusTools* (1.3). https://corpustools.readthedocs.io/en/latest/index.html

Han, Z., & Tarone, E. (2014). *Interlanguage: Forty Years Later*. John Benjamins Publishing Company.

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning*, *62*(SUPPL. 2), 79–109. https://doi.org/10.1111/j.1467-9922.2012.00707.x

Harris, T., Yang, Z., & Hardin, J. W. (2012). Modeling underdispersed count data with generalized Poisson regression. *Stata Journal*, *12*(4), 736–747. https://doi.org/10.1177/1536867x1201200412

Hartig, F. (2020). *What does it mean if a DHARMa test is significant? #212*. https://github.com/florianhartig/DHARMa/issues/212

Hartig, F. (2021a). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models*. http://web.archive.org/web/20210528100353/https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

Hartig, F. (2021b). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models*. https://cran.r-project.org/package=DHARMa

Heeringa, W., Golubovic, J., Gooskens, C., Schüppert, A., Swarte, F., & Voigt, S. (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In C. Gooskens & R. van Bezooijen (Eds.), *Phonetics in Europe: Perception and Production* (pp. 99–137). Peter Lang. https://doi.org/10.3726/978-3-653-03517-9/8

Heeringa, W., & Prokić, J. (2018). Computational dialectology. In C. Boberg, J. Nerbonne, & W. Dominic (Eds.), *The handbook of dialectology* (pp. 330–347). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118827628.ch19

Helms-Park, R., & Dronjic, V. (2013). Crosslinguistic lexical influence: Cognate facilitation. In R. A. Alonso (Ed.), *Crosslinguistic influence in second language acquisition* (pp. 71–92). Multilingual Matters.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, *81*(396), 991–999.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008a). Advances in automated language classification. In A. Arppe, K. Sinnemäke, & U. Nikanne (Eds.), *Third workshop on quantitative investigations in theoretical linguistics (QITL3)* (pp. 40–43). University of Helsinki.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008b). Explorations in automated language classification. *Folia Linguistica*, *42*(3–4), 331–353.

Honnibal, M., Montani, I., Landeghem, S. Van, & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* (3.2.4). Python library. https://doi.org/10.5281/zenodo.1212303

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, *106*(1), 501–511. https://doi.org/10.1016/j.cognition.2007.02.001

Hout, R. van, & Vermeer, A. (2010). Comparing measures of lexical richness. In *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Benjamins. https://doi.org/10.1017/cbo9780511667268.008

Hox, J. J., Moerbeek, M., & Schoot, R. van de. (2018). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.1198/jasa.2003.s281

Huang, L. S. (2009). The potential influence of L1 (Chinese) on L2 (English) communication. *ELT Journal*, *64*(2), 155–164. https://doi.org/10.1093/elt/ccp039

Huang, Y., Geertzen, J., Baker, R., Korhonen, A., & Alexopoulou, T. (2017). *The EF-Cambridge open language database (EFCAMDAT): Information for users* (pp. 1–18). https://corpus.mml.cam.ac.uk/

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*(1), 28–54. https://doi.org/10.1075/ijcl.16080.hua

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2020). Subcategorization frame identification for learner English. *International Journal of Corpus Linguistics*, *Advanced o*. https://doi.org/10.1075/ijcl.18097.hua

Ipek, H. (2009). Comparing and Contrasting First and Second Language Acquisition: Implications for Language Teachers. *English Language Teaching*, *2*(2), 155–163. https://doi.org/10.5901/mjss.2013.v4n10p56

Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12752–12757. https://doi.org/10.1073/pnas.1500331112

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, *5*, 1–16. https://doi.org/10.1038/sdata.2018.189

James, M. A. (2012). Cross-Linguistic Influence and Transfer of Learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 858–861). Springer.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning*, *50*(2), 245–309. https://doi.org/10.1111/0023-8333.00118

Jarvis, S. (2009). Lexical transfer. In A. Pavlenko (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 99–124). Multilingual Matters.

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, *63*(SUPPL. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2015). The scope of transfer research. In L. Yu & T. Odlin (Eds.), *New Perspectives on Transfer in Second Language Learning*. Multilingual Matters.

Jarvis, S. (2017). Transfer: An overview with an expanded scope. In A. Golden, S. Jarvis, & K. Tenfjord (Eds.), *Crosslinguistic influence and distinctive patterns of language learning* (pp. 12–28). Multilingual Matters. https://doi.org/10.21832/GOLDEN8767

Jarvis, S., Castañeda-Jiménez, G., & Nielsen, R. (2012). Detecting L2 writers' L1s on the basis of their lexical styles. In S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 34–70). Multilingual Matters. https://doi.org/10.21832/9781847696991-003

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Jiang, N. (1999). Testing processing explanations for the asymmetry in masked cross-language priming. *Bilingualism: Language and Cognition*, *2*(1), 59–75. https://doi.org/10.1017/S1366728999000152

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47–77. https://doi.org/10.1093/applin/21.1.47

Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*, 617–637. https://doi.org/10.1017/S0272263102004047

Jiang, N. (2004). Semantic Transfer and Its Implications for Vocabulary Teaching in a Second Language. *The Modern Language Journal*, *88*(3), 416–432. https://doi.org/10.1111/j.0026-7902.2004.00238.x

Jiang, X., Guo, Y., Geertzen, J., Alexopoulou, D., Sun, L., & Korhonen, A. (2014). Native language identification using large, longitudinal data. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 3309–3312). European Language Resources Association.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, *37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In *Probability Theory in Linguistics* (pp. 39–96). The MIT Press.

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling Methods for Identifying Outliers. *International Journal of Statistics and Systems*, *10*(2), 231–238.

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112–134). Newbury House.

Kellerman, E. (1995). Crosslinguistic influence: Transfer to nowhere? *Annual Review of Applied Linguistics*, *15*(1995), 125–150.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 60–66. https://aclanthology.org/E95-1009

Key, M. R., & Comrie, B. (2015). *The intercontinental dictionary series*. Max Planck Institute for Evolutionary Anthropology. https://ids.clld.org/

Kida, S., & Barcroft, J. (2018). Semantic and Structural Tasks for the Mapping Component of L2 Vocabulary Learning: Testing the Topra Model From a New Angle. *Studies in Second Language Acquisition*, *40*, 1–26. https://doi.org/10.1017/S0272263117000146

Koda, K. (1996). L2 word recognition research. *The Modern Language Journal*, *80*(4), 450–460.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *01*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 522–532. https://doi.org/10.1016/j.system.2012.10.017

Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. *System*, *42*(1), 23–33. https://doi.org/10.1016/j.system.2013.10.019

Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000)*, 288–295. http://dl.acm.org/citation.cfm?id=974343

Krashen, S. D. (1989). We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal*, *73*(4), 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.x

Krashen, S. D. (2003). *Explorations in Language Acquisition and Use: The Taipei Lectures*. Heinemann.

Kroll, J. F., Dussias, P. E., Bogulski, C. A., & Kroff, J. R. V. (2012). Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 56, pp. 229–262). Academic Press. https://doi.org/10.1016/B978-0-12-394393-4.00007-8

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections between Bilingual Memory Representations. *Journal of Memory and Language*, *33*, 149–174.

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism*, *13*(3), 373–381. https://doi.org/10.1017/S136672891000009X

Kubota, R. (1998). An investigation of L1–L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing*, *7*(1), 69–100. https://doi.org/10.1016/S1060-3743(98)90006-6

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, *33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Kumpf, L. (1984). Temporal systems and universality in interlanguage: A case study. In F. R. Eckman (Ed.), *Universals of Second Language Acquisition* (pp. 132–143). Newbury

House Publishers.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Kyle, K. (2018). *lexical-diversity* (0.1.0). Python library. https://github.com/kristopherkyle

Kyle, K. (2022). *pylats* (0.24). Python library. https://pypi.org/project/pylats/

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S. A., & Kim, Y. (2015). Native language identification and writing proficiency. *International Journal of Learner Corpus Research*, *1*(2), 187–209. https://doi.org/10.1075/ijlcr.1.2.01kyl

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340. https://doi.org/10.1177/0265532215587391

Lang, D. T. (2020). *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package. https://cran.r-project.org/package=XML

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production Willem. *Behavioral and Brain Sciences*, *22*, 1–75.

Levshina, N. (2018). Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help + (to) Infinitive in varieties of web-based English. *Glossa: A Journal of General Linguistics*, *3*(1), 1–22. https://doi.org/10.5334/gjgl.294

Lim, J. M.-H. (2007). Crosslinguistic influence versus intralingual interference: A pedagogically motivated investigation into the acquisition of the present perfect. *System*, *35*(3), 368–387. https://doi.org/10.1016/j.system.2007.03.002

Lindgren, J., & Bohnacker, U. (2020). Vocabulary development in closely-related languages: Age, word type and cognate facilitation effects in bilingual Swedish-German preschool children. *Linguistic Approaches to Bilingualism*, *10*(5), 587–622. https://doi.org/10.1075/lab.18041.lin

Little, C. C. (2018). *Abydos NLP/IR library for Python* (0.3.5). http://doi.org/10.5281/zenodo.1463204

Llach, M. P. A. (2010). An overview of variables affecting lexical transfer in writing: A review study. *International Journal of Linguistics*, *2*(1), E2. https://doi.org/10.5296/ijl.v2i1.445

Llach, M. P. A., Fontecha, A. F., & Espinosa, S. M. (2006). Differences in the written production of young Spanish and German learners: evidence from lexical errors in a composition. *Barcelona English Language and Literature Studies*, *14*, 2–13.

Lotto, L., & de Groot, A. M. B. (1998). Effects of Learning Method and Word Type on Acquiring Vocabulary in an Unfamiliar Language. *Language Learning*, *48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

Lüdecke, D., Ben-shachar, M. S., Patil, I., Makowski, D., Waggoner, P., Patil, I., Ben-shachar, M. S., Patil, I., & Makowski, D. (2021). Assessment of regression models performance.

The Journal of Open Source Software, *6*(59), 1–8. https://doi.org/10.21105/joss.03132

Luján-Mora, S., & Palomar, M. (2001). Comparing string similarity measures for reducing inconsistency in integrating data from different sources. *Proceedings of the Second International Conference on Advances in Web-Age Information Management (WAIM 2001)*, *2118*, 191–202. https://doi.org/10.1007/3-540-47714-4_18

Lynch, H. J., Thorson, J. T., & Shelton, A. O. (2014). Dealing with under- and over-dispersed count data in life history, spatial, and community ecology. *Ecology*, *95*(11), 3173–3180.

Major, R. C. (1998). Interlanguage Phonetics and Phonology. *Studies in Second Language Acquisition*, *20*(02), 131–137. https://doi.org/10.1017/S0272263198002010

Makowski, D., & Lüdecke, D. (2019). *The report package for R: Ensuring the use of best practices for results reporting*. R package. https://github.com/easystats/report

Malmasi, S., & Dras, M. (2015). Large-scale Native Language Identification with cross-corpus evaluation. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 1403–1409. https://doi.org/10.3115/v1/n15-1160

Manurung, R., Ritchie, G., Pain, H., Waller, A., Black, R., & O’Mara, D. (2008). Adding phonetic similarity data to a lexical database. *Language Resources and Evaluation*, *42*(3), 319–324. https://doi.org/10.1007/s10579-008-9069-5

Marecka, M., Szewczyk, J., Otwinowska, A., Durlik, J., Foryś-Nogala, M., Kutyłowska, K., & Wodniecka, Z. (2021). False friends or real friends? False cognates show advantage in word form learning. *Cognition*, *206*, 104477. https://doi.org/10.1016/j.cognition.2020.104477

Martin, K. I., & Ellis, N. C. (2012). The Roles of Phonological Short-Term Memory and Working Memory in L2 Grammar and Vocabulary Learning. *Studies in Second Language Acquisition*, *34*(03), 379–413. https://doi.org/10.1017/S0272263112000125

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McCallum, J. (2019). *autocorrect* (0.4.4). Python library. https://github.com/phatpiglet/autocorrect/

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCoy, R. T., & Frank, R. (2018). Phonologically Informed Edit Distance Algorithms for Word Alignment with Low-Resource Languages. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, *1998*, 102–112. https://doi.org/10.7275/R5251GC0

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, *39*, 74–92. https://doi.org/10.1017/s0267190519000096

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.

http://conference.scipy.org/proceedings/scipy2010/mckinney.html

McSweeny, J. L., & Shriberg, L. D. (1995). Segmental and suprasegmental transcription reliability. In *Phonology Project Technical Report No. 2*.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*(March 2019), 104092. https://doi.org/10.1016/j.jml.2020.104092

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Routledge.

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, *3*(2), 124–152.

Morrissey, M. B., & Ruxton, G. D. (2018). Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, *10*(3), 1–24. https://doi.org/10.3998/ptpbio.16039257.0010.003

Mortensen, D. R. (2015). *PanPhon*. https://github.com/dmort27/panphon

Mortensen, D. R. (2019). *Epitran*. http://archive.is/2020/https://github.com/dmort27/epitran

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2710–2714.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3475–3484.

Muljani, D., Koda, K., & Moates, D. R. (1998). Development of word recognition in a second language. *Applied Psycholinguistics*, *19*, 99–113. https://doi.org/10.4324/9781315882741

Murakami, A. (2014). *Individual variation and the role of L1 in the L2 development of English grammatical morphemes: insights from learner corpora* [University of Cambridge]. https://doi.org/https://doi.org/10.17863/CAM.16509

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *66*(4), 834–871. https://doi.org/10.1111/lang.12166

Murakami, A., & Alexopoulou, T. (2016). L1 Influence on the Acquisition Order of English Grammatical Morphemes: A Learner Corpus Study. *Studies in Second Language Acquisition*, *38*, 365–401. https://doi.org/10.1017/S0272263115000352

Murphy, S. (2003). Second language transfer during third language acquisition. *Working Papers in TESOL & Applied Linguistics*, *3*(1), 1–21.

Nerbonne, J., & Heeringa, W. (1997). Measuring dialect distance phonetically. In J. Coleman (Ed.), *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics* (Issue 1995, pp. 11–18).

Nerbonne, J., & Heeringa, W. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, *9*, 69–84. https://doi.org/10.1515/dig.2001.2001.9.69

Nisioi, S. (2015). Feature Analysis for Native Language Identificatio. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science* (Vol. 9041, pp. 644–657). Springer International Publishing. https://doi.org/10.1007/978-3-319-18111-0

O'Brien, I. (1998). Phonological Memory Predicts Second Language Oral Gains in Adults. *Psychological Review,* 8888–8888.

O'Brien, I., Segalowitz, N., Collentine, J. O. E., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*(3), 377–402. https://doi.org/10.1017.S0142716406060322

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

O'Sullivan, Í., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, *15*(1), 49–68. https://doi.org/10.1016/j.jslw.2006.01.002

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Odlin, T. (2003). Cross-Linguistic Influence. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 436–485). Blackwell Publishing.

Odlin, T. (2005). Crosslinguistic influence and conceptual transfer: What are the concepts? *Annual Review of Applied Linguistics*, *25*, 3–25.

Odlin, T. (2013). Crosslinguistic influence in second language acquisition. In C. A. Chapelle (Ed.), *The ecyclopedia of applied linguistics* (pp. 1562–1568). Blackwell Publishing. https://doi.org/10.1002/9781405198431.wbeal0292

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol Publishing.

Ooms, J. (2018). *cld2: Google's Compact Language Detector 2*. R package. https://cran.r-project.org/package=cld2

Ota, M., & Green, S. J. (2013). Input frequency and lexical variability in phonological development: a survival analysis of word-initial cluster production. *Journal of Child Language*, *40*(3), 539–566. https://doi.org/10.1017/S0305000912000074

Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-linguistic similarity matter. *Language Learning*, *70*(3), 685–731. https://doi.org/10.1111/lang.12390

Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, *22*(8), 974–991. https://doi.org/10.1080/13670050.2017.1325834

Petroni, F., & Serva, M. (2008). Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(8). https://doi.org/10.1088/1742-5468/2008/08/P08012

Petroni, F., & Serva, M. (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and Its Applications*, *389*(11), 2280–2283. https://doi.org/10.1016/j.physa.2010.02.004

Phillips, M. (2019). *English to IPA*. Python library. https://github.com/mphilli/English-to-IPA

Pompei, S., Loreto, V., & Tria, F. (2011). On the accuracy of language trees. *PLoS ONE, 6*(6), e20109. https://doi.org/10.1371/journal.pone.0020109

Poort, E. D., & Rodd, J. M. (2017). The cognate facilitation effect in bilingual lexical decision is influenced by stimulus list composition. *Acta Psychologica*, *180*, 52–63. https://doi.org/10.1016/j.actpsy.2017.08.008

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rabinovich, E., Tsvetkov, Y., & Wintner, S. (2018). Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, *6*, 329–342. https://doi.org/10.1162/tacl_a_00024

Ramsdell, H. L., Oller, D. K., & Ethington, C. A. (2007). Predicting phonetic transcription agreement: Insights from research in infant vocalizations. *Clinical Linguistics & Phonetics*, *21*(10), 793–831. https://doi.org/10.1080/02699200701547869

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. http://is.muni.cz/publication/884893/en

Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167–188). TESOL.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Multilingual Matters.

Ringbom, H. (1992). On L1 Transfer in L2 Comprehension and L2 Production. *Language Learning*, *42*(1), 85–112. https://doi.org/10.1111/j.1467-1770.1992.tb00701.x

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283.

Rubenstein, H., & Pollack, I. (1963). Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 147–158.

Sáchez-Casas, R. M., García-Albea, J. E., & Christopher, D. W. (1992). Bilingual lexical processing: Exploring the cognate/non-cognate distinction. *European Journal of Cognitive Psychology*, *4*(4), 293–310.

Sadat, J., Pureza, R., & Alario, F. X. (2016). Traces of an early learned second language in discontinued bilingualism. *Language Learning*, *66*(Suppl. 2), 210–233. https://doi.org/10.1111/lang.12199

Sánchez-Casas, R., & García-Albea, J. E. (2005). The Representation of Cognate and Noncognate Words in Bilingual Memory: Can Cognate Status Be Characterized as a Special Kind of Morphological Relation? Rosa. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 226–250). Oxford University Press.

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, *16*(1), 96–114. https://doi.org/10.1080/09296170802514138

Santos, R., Murrieta-Flores, P., & Martins, B. (2018). Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*, *11*(9), 913–938. https://doi.org/10.1080/17538947.2017.1371253

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism*, *15*(1), 157–166. https://doi.org/10.1017/S1366728910000623

Schepens, J., Dijkstra, T., Grootjen, F., & van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0063006

Schepens, J., van der Slik, F., & van Hout, R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning*, *66*(1), 224–256. https://doi.org/10.1111/lang.12150

Schepens, J., van der Slik, F., & van Houta, R. (2013a). Learning complex features: A morphological account of L2 learnability. *Language Dynamics and Change*, *3*(2), 218–244. https://doi.org/10.1163/22105832-13030203

Schepens, J., van der Slik, F., & van Houta, R. (2013b). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In L. Borin & A. Saxena (Eds.), *Approaches to measuring linguistic differences* (pp. 199–230). De Gruyter.

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 1–14. https://doi.org/10.1016/j.cognition.2019.104056

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, *12*(1), 40–72. https://doi.org/10.1177/026765839601200103

Sedgwick, P. (2010). Incidence rate ratio. *BMJ*, *341*, c4804. https://doi.org/10.1136/bmj.c4804

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia*, *20*(6), 615–627.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*, 209–232.

Selinker, L. (2013). *Rediscovering Interlanguage*. Routledge.

Sellers, K. F., & Morris, D. S. (2017). Underdispersion models: Models that are “under the radar.” *Communications in Statistics - Theory and Methods*, *46*(24), 12075–12086. https://doi.org/10.1080/03610926.2017.1291976

Sersen, W. J. (2011). Improving Writing Skills of Thai EFL Students by Recognition of and Compensation for Factors of L1 to L2 Negative Transfer. *US-China Education Review A*, *1*(3), 339–345.

Serva, M., & Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, *81*(6), 1–5. https://doi.org/10.1209/0295-5075/81/68005

Shatz, I. (2017). Native language influence during second language acquisition: A large-scale learner corpus analysis. In M. Hirakawa, J. Matthews, K. Otaki, N. Snape, & M. Umeda (Eds.), *Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016)* (pp. 175–180). Japan Second Language Association.

Shatz, I. (2019). How native language and L2 proficiency affect EFL learners' capitalisation abilities: A large-scale corpus study. *Corpora*, *14*(2), 173–202. https://doi.org/10.3366/cor.2019.0168

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 221–237. https://doi.org/10.1075/ijlcr.20009.sha

Sheng, L., Lam, B. P. W., Cruz, D., & Fulton, A. (2016). A robust demonstration of the cognate facilitation effect in first-language and second-language naming. *Journal of Experimental Child Psychology*, *141*, 229–238. https://doi.org/10.1016/j.jecp.2015.09.007

Shriberg, L. D., & Lof., G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, *5*(3), 225–279.

Silveira, A. P. da, & Leussen, J.-W. van. (2015). Generating a bilingual lexical corpus using interlanguage normalized Levenshtein distances. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 1–5.

Sofroniev, P. (2018). *asjp* (0.0.2). Python library. https://github.com/pavelsof/asjp

Solc, T. (2019). *Unidecode* (1.1.1). Python library. https://pypi.org/project/Unidecode/

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*(2), 293–321. https://doi.org/10.1017/S0142716404001146

Speer, R. (2020). *wordfreq 2.2.2*. PyPi. https://pypi.org/project/wordfreq/

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018). *wordfreq* (2.2). LuminosoInsight. https://doi.org/10.5281/zenodo.1443582

Stemle, E., & Onysko, A. (2015). Automated L1 identification in English learner essays and its implications for language transfer. In H. Peukert (Ed.), *Transfer effects in multilingual language development* (pp. 297–321). John Benjamins Publishing Company. https://doi.org/10.1075/hsld.4.13ste

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics*, *16*(4), 157–167.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, *21*(2), 121–137.

Talamas, A., Kroll, J. F., & Dufour, R. (1999). From form to meaning: Stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition*, *2*(1), 45–58. https://doi.org/10.1017/S1366728999000140

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96–116.

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, *104*(30), 12530–12535. https://doi.org/10.1073/pnas.0609927104

Tolentino, L. C., & Tokowicz, N. (2011). Across Languages, Space, and Time. *Studies in Second Language Acquisition*, *33*(01), 91–125. https://doi.org/10.1017/S0272263110000549

Tolentino, L. C., & Tokowicz, N. (2014). Cross-Language Similarity Modulates Grammar

Instruction. *Language Learning*, *64*(June), 279–309. https://doi.org/10.1111/lang.12048

Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, *59*(3), 623–646. https://doi.org/10.1111/j.1467-9922.2009.00519.x

Torruella, J., & Capsada, R. (2013). Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia - Social and Behavioral Sciences*, *95*, 447–454. https://doi.org/10.1016/j.sbspro.2013.10.668

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures* (pp. 79–103). John Benjamins Publishing Company. http://centaur.reading.ac.uk/28712/1/Chapter 6 (Treffers-Daller)final_25June2012.docx

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Tukey, J. W. (1977). *Exploratory Data Analysis* (First edit). Addison-Wesley.

Upton, T. A., & Lee-Thompson, L.-C. (1987). The role of the first language in second language learning. *Studies in Second Language Acquisition*, *23*, 469–495.

Urlacher, J. (2010). *Second language proficiency : Self-report vs. objective measures and relationship with sentential priming in the processing of interlingual homographs* [(M.A. Thesis). Electronic Theses and Dissertations. Paper 44.]. http://scholar.uwindsor.ca/etd/44

van de Ven, M., Segers, E., & Verhoeven, L. (2019). Enhanced second language vocabulary learning through phonological specificity training in adolescents. *Language Learning*, *69*(1), 222–250. https://doi.org/10.1111/lang.12330

Van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122. https://cran.r-project.org/package=stringdist

van der Slik, F. W. P. (2010). Acquisition of Dutch as a second language: The explanative power of cognate and genetic linguistic distance measures for 11 West European first languages. *Studies in Second Language Acquisition*, *32*(3), 401–432. https://doi.org/10.1017/S0272263110000021

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vandenberghe, B., Perez, M. M., Reynvoet, B., & Desmet, P. (2021). Combining explicit and sensitive indices for measuring L2 vocabulary learning through contextualized input and word-focused instruction. *Studies in Second Language Acquisition*, *43*(5), 1009–1039. https://doi.org/10.1017/S0272263120000431

Vanlangendonck, F., Peeters, D., Rueschemeyer, S. A., & Dijkstra, T. (2020). Mixing the stimulus list in bilingual lexical decision turns cognate facilitation effects into mirrored inhibition effects. *Bilingualism*, *23*(4), 836–844. https://doi.org/10.1017/S1366728919000531

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568–

587. https://doi.org/10.1177/1362168818817945

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Contributors, S. 1. 0. (2019). SciPy 1.0—Fundamental algorithms for scientific computing in Python. *ArXiv E-Prints*, 1–22. http://arxiv.org/abs/1907.10121

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, *13*, 22–30. https://doi.org/10.1109/MCSE.2011.37

What is the best way to remove accents in a Python unicode string? (2009). Stackoverflow. https://stackoverflow.com/questions/517923/what-is-the-best-way-to-remove-accents-in-a-python-unicode-string

Wichmann, S. (2019). How to distinguish languages and dialects. *Computational Linguistics*, *45*(4), 823–831. https://doi.org/10.1162/COLIa00366

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and Its Applications*, *389*(17), 3632–3639. https://doi.org/10.1016/j.physa.2010.05.011

Wichmann, S., Holman, E. W., & Brown, C. H. (2018). *The ASJP database* (No. 18). https://asjp.clld.org/

Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*(2), 177–197. https://doi.org/10.1515/LITY.2011.013

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2019). *dplyr: A Grammar of Data Manipulation*. R package. https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., & RStudio. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package. https://cran.r-project.org/web/packages/stringr/index.html

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. (2014). Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change*, *4*(2), 253–269.

Wieling, M., Heeringa, W., & Nerbonne, J. (2007). An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-Van Reenen-Project Data. *Taal En Tongval*, *59*, 84–116.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., & Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PLoS ONE*, *9*(1), e75734. https://doi.org/10.1371/journal.pone.0075734

Williams, J. N. (2015). The bilingual lexicon. In J. Taylor (Ed.), *The Oxford handbook of the word*. Oxford University Press.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge. https://doi.org/10.4324/9781315165547

Wu, Y. J., & Thierry, G. (2010). Chinese-English Bilinguals Reading English Hear Chinese. *Journal of Neuroscience, 30*(22), 7646–7651. https://doi.org/10.1523/JNEUROSCI.1602-10.2010

Xia, C. M. (2017). Psychotypology of Chinese learners of English and its influence on the acquisition of metaphorical expressions: An offline study. *Cambridge Occasional Papers in Linguistics*, *10*, 237–255.

Yan, X., Kim, H. R., & Kim, J. Y. (2020). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing*. https://doi.org/10.1177/0265532220951508

Yi, A. (2012). On the Factors Influencing L1 Transfer. *Theory and Practice in Language Studies, 2*(11), 2372–2377. https://doi.org/10.4304/tpls.2.11.2372-2377

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Yuan, H.-C. (2014). A Corpus-based Study on the Influence of L1 on EFL Learners' Use of Prepositions. *Theory and Practice in Language Studies, 4*(12), 2513–2521. https://doi.org/10.4304/tpls.4.12.2513-2521

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhang, J., Wu, C., Zhou, T., & Meng, Y. (2019). Cognate facilitation priming effect is modulated by writing system: Evidence from Chinese-English bilinguals. *International Journal of Bilingualism*, *23*(2), 553–566. https://doi.org/10.1177/1367006917749062

Zhang, L. (2018). *A More Sensitive Edit-Distance for Measuring Pronunciation Distances and Detecting Loanwords*. University of Groningen.

Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications*, *389*(1), 58–71. https://doi.org/10.1016/j.jmaa.2011.11.042

Zhu, Y., & Mok, P. P. K. (2020). Visual recognition of cognates and interlingual homographs in two non-native languages. *Linguistic Approaches to Bilingualism*, *10*(4), 441–470. https://doi.org/10.1075/lab.17049.zhu

7 APPENDICES

7.1 Appendix A: Further background information

7.1.1 Examples of lexical transfer

One example of lexical transfer appears in a study by N. Jiang (2004), who showed that semantic transfer from learners' L1 leads them to process pairs of L2 words faster when they have the same L1 translation compared to when they do not. This means that, as Jiang points out (p. 421), when Korean learners of English as a second language (ESL) are shown pairs of words in English, they are able to process those pairs faster when the words are similar in meaning and share the same Korean translation, compared to when the words are similar in meaning but have different translations. For instance, Korean speakers are faster when they process word pairs such as *chance* and *opportunity*, which are similar in meaning and which share the same Korean translation (*기회*, pronounced 'giwhoi'), than when they process word pairs such as *decrease* and *reduce*, which are also similar in meaning but which have different Korean translations (*줄이다*, which is pronounced 'julida', and 축소되다, which is pronounced 'chuksohada').[41] Jiang attributed this phenomenon to new L2 words being linked to their L1 translation, which results in information from the L1 lemma transferring to the L2 lexemes.[42]

Another example of lexical transfer appears in a study by Tonzar, Lotto, and Job (2009), who examined how Italian-speaking children learned L2 words in English and German. They found that when a word in the children's target L2 was part of a cognate pair with an L1 word (e.g., *limone-lemon*, but not *mela-apple*), then the children had an easier time acquiring it. However, they also found that this effect decreased over time, as children's familiarity with the target L2 grew.

Furthermore, Tonzar et al. (2009) also found that similarities and differences in form between the children's L1 and their target L2 influenced the L2 errors that they made. Specifically, when the L2 words were part of cognate pair with the L1, learners tended to make mistakes that maximized the *similarities* between the L2 and the L1, using one of two possible strategies:[43]

[41] These examples are all taken directly from Jiang's paper (2004, p.421).

[42] In this context, the term *lemma* refers to the part of the word that contains its syntactic and semantic information (N. Jiang, 2000). However, in some works, the term *lemma* is used to refer to a word's syntax, but not to its semantics, which is viewed as its *meaning*, or its underlying lexical concept (Levelt et al., 1999).

[43] All the examples that are described here are taken from the paper by Tonzar et al. (2009, p, 640).

i. Transforming L2 phonemes into L1 graphemes. For example, writing *carafe* in German instead of *karaffe* (“carafe”), since the /k/ sound is spelled as /c/ in Italian.
ii. Adapting L2 words to L1 word forms. For example, writing *polipe* in German instead of *polyp* (“polyp”), since nouns in Italian generally end with a vowel.

Conversely, when the L2 words were non-cognates, learners tended to make mistakes that maximized the *differences* between the L2 and the L1, using one of two possible strategies:

i. Adapting L1 words to L2 word forms. For example, adding a word-final consonant in order to differentiate the L2 word from L1 (Italian) words that generally have a word-final vowel, as in the case of writing *mühler* in German instead of *mühle* (“mill”).
ii. Using L2 graphemes that are not a part of the L1 alphabet. For example, inserting a ‘ß’, as in the case of writing *flatoß* in German instead of *flöte* (“flute”).

7.1.2 Theories of lexical acquisition and transfer

[Here, I outline a few theories that can explain various aspects of lexical acquisition and transfer, in the context of the L2 lexicon. This is meant to provide some insights into the topic, but is *not* meant to be a comprehensive review. Furthermore, my research does not make a claim regarding which of these particular theories can best explain my findings, because the focus of the research was different, and as such it was not conducted in a way that would allow me to answer this question.]

One theory of lexical acquisition and transfer is N. Jiang's (2000, 2002, 2004) psycholinguistic model. In this model, development of new L2 lexical entries occurs in three stages. The first is the *formal stage*, when an initial lexical entry is established; this entry consists of formal phonological and orthographical specifications, together with a pointer to the L1 translation. The second stage is the *lemma-mediation stage*, when information from the L1 lemma is transferred to the related L2 lexical items, and mediates their usage. The final stage is the *L2-integration stage*, when syntactic, semantic, and morphological specifications from the L2 are fully integrated into the L2 lexical entries.

Under this model, it is expected that the majority of L2 words will generally fossilize at the second stage, and these fossilized entries consist of formal L2 specifications, together with the syntactic and semantic information that was transferred from their L1 counterparts.[44]

[44] N. Jiang (2000) points out two main factors that encourage L2 learners to rely on their L1 during L2 lexical development. First, unlike during L1 acquisition, during L2 acquisition learners already have a preexisting

For example, consider the following sentences (taken from N. Jiang, 2002, p. 620), where the incorrectly used fossilized words are underlined, while the felicitous words (which should be used in this context) are provided in parenthesis:

i. I go to the oven (bakery) in the morning to buy bread.

ii. He bit himself in the language (tongue).

The sentence in (i) was produced by a native speaker of Arabic, where both the meaning of *oven* and the meaning of *bakery* are linked to the same word: *furn*. Similarly, the sentence in (ii) was produced by a native Finnish speaker, where both the meaning of *language* and the meaning of *tongue* are also linked to the same word: *kieli*. In both cases, the cause of the error is the same: there is a certain distinction that exists between two meanings in the L2, but not in the L1, so the two L2 words are mapped to the same L1 translation. This causes learners to struggle to distinguish between the L2 words from a semantic perspective, which makes it difficult for them to choose the correct word during L2 production (Jarvis & Pavlenko, 2008; N. Jiang, 2002).[45] Furthermore, such errors have been found to occur consistently even for high-frequency L2 words, and even among learners with high L2 proficiency, which provides support for the notion of fossilization (Altenberg & Granger, 2001).

Other theories of L2 lexical development predict a similar influence of learners' L1 on their acquisition of L2 lexical items. For example, the *Parasitic Model* of vocabulary acquisition suggests that L2 entries are initially mapped to an L1 lemma, since learners exploit existing lexical material in their L1 in order to establish an initial representation of words in the L2 (Ecke, 2015; C. J. Hall, 2002; C. J. Hall & Ecke, 2003). Similarly the *Revised Hierarchical Model* suggests that during the initial stages of acquisition, learners access L2 words through the L1, and this form of mediated conceptual processing of L2 words is influenced by form associations between L2 words and their L1 counterparts (Kroll et al., 2010; Kroll & Stewart, 1994; Talamas et al., 1999).

These theories of L2 lexical development therefore suggest that formal similarity between words in learners' L1 and their equivalents in the target L2 can facilitate the acquisition of the L2 words, by helping learners form a connection between the L1 and L2 lexical items, which makes it easier to use relevant lexical information from the L1 during the

conceptual/semantic system, which is associated with their L1 lexical system. Second, L2 acquisition occurs primarily through instructed learning, which often involves poor input in terms of quality and quantity.

[45] This type of error involves *semantic* transfer, but similar errors can also involve *conceptual* transfer, depending on the words and concepts involved (Jarvis & Pavlenko, 2008, pp. 120–122).

acquisition of the L2 items. However, there are also additional factors that can explain why various types of similarity between learners' L1 and their target L2 can facilitate acquisition.

One notable reason why similarity in form between L1 words and their L2 counterparts could facilitate the acquisition of these L2 words is that such similarity could reduce the cognitive processing load that is required both for the L2-L1 mapping and for the learning of the structural properties of the L2 words. This, in turn, could allow the learners to dedicate more cognitive resources to the acquisition of other aspects of the words, and most notably its semantic properties, such as its appropriate usage space or its collocational properties (Kida & Barcroft, 2018). This is reflected in the *Type of Processing–Resource Allocation* (TOPRA) model, which suggests that *increasing* the processing demands in one domain, such as structure, could *decrease* learners' ability to process other domains, such as semantics (Barcroft, 2004; Kida & Barcroft, 2018). Accordingly, similarity in structure between L1 words and their L2 counterparts could have an opposite effect, since a *decrease* in the structural processing demands of the L2 words and in the processing demands of the L2-L1 mapping, could allow for an *increase* in learners' ability to process the semantics of those words.[46]

Furthermore, the idea that decreasing the processing demands in one domain could allow learners to dedicate more cognitive resources to other domains could also explain why L1-L2 lexical similarity could facilitate the acquisition of lexical items that are not a part of a cognate pair. Specifically, the fact that learners have to dedicate less time and resources when learning L2 cognates, as a result of the facilitative effect of L1-L2 similarity, this means that they can dedicate more time and mental resources to learning words that are not a part of a cognate pair.

Finally, there are additional factors which could be responsible for the facilitative effect of L1-L2 lexical similarity in cognate pairs. For examples, it is possible that the structural similarity serves as a beneficial cue for the retrieval of the corresponding translation (Tonzar et al., 2009). In addition, some researchers propose that while non-cognates are represented through separate entries in the learner's lexicon, cognates share a single entry, which is connected to a single underlying concept, which could facilitate acquisition of cognates, as it is easier for learners to modify an existing lexical entry instead of creating a new one (Sáchez-Casas et al., 1992; Sánchez-Casas & García-Albea, 2005; Tonzar et al., 2009).

[46] In this context, L2-L1 mapping refers to a process where a new L2 word form is mapped to a known L1 meaning, as L2 learners are generally expected to transfer semantic representations from the L1, rather than relearn the meaning of words such as *door* or *apple* (Kida & Barcroft, 2018).

7.1.3 False cognates

Though crosslinguistic lexical similarity often leads to positive lexical transfer, it can also lead to negative transfer (i.e., interference). A notable example of this is the case of *false cognates* (sometimes also referred to as *false friends*, *deceptive cognates*, *homographic non-cognates*, *interlexical homographs/homophones*, and *interlingual homographs/homophones*), which are crosslinguistic word pairs that have similar form in terms of phonology and/or orthography, but different meanings (primarily in terms of semantics), unlike cognates, which are similar in both form and meaning (Brenders et al., 2011; Chavula & Suleman, 2016; Gerard & Scarborough, 1989; Marecka et al., 2021; Otwinowska & Szewczyk, 2019; Ringbom, 2007; Schepens, Dijkstra, et al., 2013; Urlacher, 2010).[47] For example, in the case of English and Dutch, the word *room* is a false cognate, since it means 'room' in English and 'cream' in Dutch (Chavula & Suleman, 2016).

Unlike regular cognates, which learners generally find easier to process, learn, and use than non-cognates, learners generally struggle when it comes to engaging with false cognates (Otwinowska & Szewczyk, 2019; Tonzar et al., 2009). This is especially an issue when there is some semantic overlap between the L1 and the L2 words, or if they frequently occur in the same context, since this can increase interference from the L1 (Ringbom, 2007). An example of this is presented by Ringbom (2007), who points out the English-Swedish adjective pair of *phoney* and *fånig*, which have a similar form, a similar negative connotation, and which are used in similar situations, but which also have slightly different meanings, as the English word means that something is 'fake', while the Swedish word means that something is 'silly' or 'ridiculous'. However, research suggests that false cognates are generally relatively rare compared to cognates (Chavula & Suleman, 2016; Ringbom, 2007), and in the case of French and English, for example, estimates suggest that there is approximately 1 false cognate for every 11 cognates (Ringbom, 2007).

[47] This distinction between cognates and false cognates is based on the common distinction used in the context of psycholinguistics. Other distinctions may also be used, particularly in the context of historical linguistics when taking the etymology of the words into account (S. E. Carroll, 1992).

7.2 Appendix B: Sample information

This document contains background information about the EFCAMDAT and the EFCAMDAT Cleaned Subcorpus, as well as about the sample selection process and the final sample that was used in the study.

7.2.1 Background information on the EFCAMDAT

The *EF Cambridge Open Language Database* (EFCAMDAT) is an open access L2 English/EFL learner corpus, available at https://corpus.mml.cam.ac.uk/. It has been extensively used to study L2 acquisition, including when it comes to task effects (Alexopoulou et al., 2017; Michel et al., 2019) and L1 effects (Alexopoulou et al., 2015; Chen et al., 2021; X. Jiang et al., 2014; Murakami, 2014, 2016; Shatz, 2017, 2019).[48]

The EFCAMDAT contains over 1,180,00 texts, written by approximately 175,000 learners from various nationalities, who were enrolled in Education First's (EF) online English school, called "Englishtown", between 2011–2013 (Geertzen et al., 2013; Y. Huang et al., 2017, 2018). When a learner joins EF's online school, they are given an English proficiency placement test, and are allocated a starting proficiency level accordingly (Geertzen et al., 2013). The EFCAMDAT spans 16 teaching levels of increasing proficiency, which EF has aligned with common proficiency standards (Geertzen et al., 2013), such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001).[49] Specifically:

> EF teaching materials are designed to use tasks that align with the can-do statements for the relevant CEFR level, as well as vocabulary and grammatical structures that are appropriate for the level. These alignments are based on the Council of Europe's CEFR documentation, criterial feature research, and the content developers' experience.
>
> (Alexopoulou et al., 2017, appendix S1, p.1)

Each level consists of 6–8 distinct lessons. After completing each lesson, learners are assigned a short writing task that they submit online. As Alexopoulou et al. (2017, p. 192) note:

[48] Murakami (2013) also compared some L1 effects that were found in the EFCAMDAT with those found in the Cambridge Learner Corpus (CLC), and found that there was a strong correlation between the two.

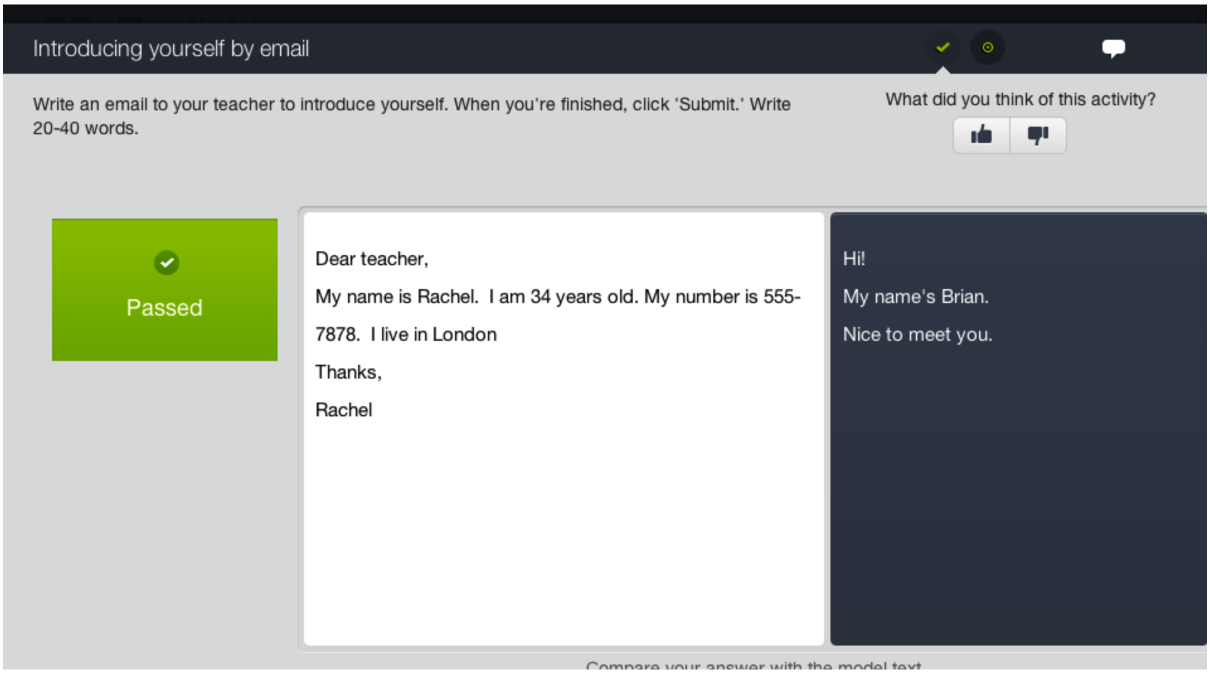
[49] EF's proficiency classification scheme has been used extensively in previous studies that used the dataset (e.g. Alexopoulou et al., 2015, 2017; Chen et al., 2020; Michel et al., 2019; Murakami, 2016; Shatz, 2019), and support for this scheme also comes from past studies, such as Murakami (2013), who analyzed both the EFCAMDAT and the Cambridge Learner Corpus (CLC), and concluded that the correspondence between the EF levels and the CEFR levels seems to hold, at least for the A2 and B1 levels, which are the levels that he was able to directly compare to their equivalent levels in the CLC (p. 84).

> Broadly speaking, the writing activities contained in EFCAMDAT can be characterized as tasks, because L2 writers work toward a nonlinguistic outcome, for example, write a complaint or apply for a job, and are engaged in language use to achieve that goal (Samuda & Bygate, 2008). In this sense, EFCAMDAT can be seen as a task-based corpus.

Tasks include a prompt and general instructions on what to write and how long to make the text, and have an expected length, ranging from 20–40 words in lesson 1 in level 1, to 150–180 words in lesson 8 in level 16 (Alexopoulou et al., 2017). The tasks cover a wide range of topics, such as describing your favorite day or reviewing a song for a website, and can involve different task types, such as *narrative* and *descriptive*. For an overview of how tasks in the EFCAMDAT may be classified, see Alexopoulou et al. (2017) and Michel et al. (2019).

Figures 12 and 13 contain examples of the screens learners see while performing the writing tasks. Table 12 contains the prompt of the first task in each level. A list of all task prompts, together with screenshots of the tasks up to and including level 10, can be found on the EFCAMDAT website (https://corpus.mml.cam.ac.uk/task_screenshots/index.php).[50]

[50] We include a representative sample is of these prompts and screenshots—rather than all of them—for copyright and length reasons. However, as noted, all this information can be accessed on the EFCAMDAT site.


What did you think of this activity?
Passed
Compare your answer with the model text

Introducing yourself by email

Write an email to your teacher to introduce yourself. When you're finished, click 'Submit.' Write 20-40 words.

Dear teacher,

My name is Rachel. I am 34 years old. My number is 555-7878. I live in London

Thanks,

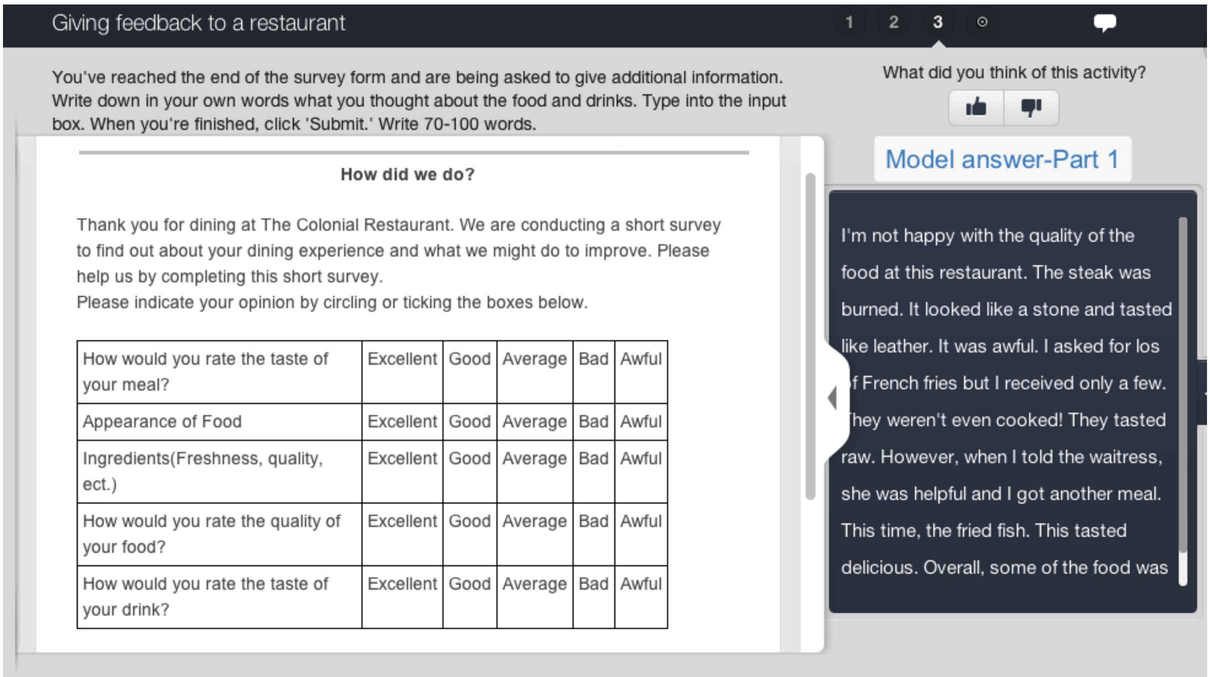
Rachel

Hi!

My name's Brian.

Nice to meet you.

Figure 12. A screenshot for the first task in level 1, about "Introducing yourself by email". At the top of the screen is the prompt; learners write the text in the box in the middle. On the right is a model answer that learners can choose to view.



Giving feedback to a restaurant

You've reached the end of the survey form and are being asked to give additional information. Write down in your own words what you thought about the food and drinks. Type into the input box. When you're finished, click 'Submit.' Write 70-100 words.

How did we do?

Thank you for dining at The Colonial Restaurant. We are conducting a short survey to find out about your dining experience and what we might do to improve. Please help us by completing this short survey.
Please indicate your opinion by circling or ticking the boxes below.

How would you rate the taste of your meal?	Excellent	Good	Average	Bad	Awful
Appearance of Food	Excellent	Good	Average	Bad	Awful
Ingredients(Freshness, quality, ect.)	Excellent	Good	Average	Bad	Awful
How would you rate the quality of your food?	Excellent	Good	Average	Bad	Awful
How would you rate the taste of your drink?	Excellent	Good	Average	Bad	Awful

I'm not happy with the quality of the food at this restaurant. The steak was burned. It looked like a stone and tasted like leather. It was awful. I asked for los f French fries but I received only a few. hey weren't even cooked! They tasted raw. However, when I told the waitress, she was helpful and I got another meal. This time, the fried fish. This tasted delicious. Overall, some of the food was

Figure 13. A screenshot for the first task in level 9, about "Giving feedback to a restaurant". At the top of the screen is the prompt, followed by relevant background content in the middle that learners write their text in response to. On the right is a model answer that learners can choose to view.

Table 12. Examples of the topics and prompts of the first task in each level in the EFCAMDAT. *Task* is the task number out of all the tasks (1–128). *Level* is the EFCAMDAT level (1–16). *CEFR* is the corresponding CEFR level (there are 3 EFCAMDAT levels per CEFR level in all cases, except for the final CEFR level C2, which corresponds to EFCAMDAT level 16 only). *Topic* is the topic of the task, and *Prompt* is the prompt that learners are given for the task.

Task	Level	CEFR	Topic	Prompt
1	1	A1	Introducing yourself by email	Write an email to your teacher to introduce yourself. When you're finished, click 'Submit.' Write 20-40 words.
9	2	A1	Describing your favorite day	What's your favorite day of the week? What do you usually do on that day, and at what time? Write about your favorite day of the week below. Remember to include an introduction, middle and end to your writing. Type into the input box. When you're finished, click 'Submit.' Write 20-40 words.
17	3	A1	Replying to a new penpal	You receive this email from a new online pen friend. Write an appropriate reply. Type into the input box. When you're finished, click 'Submit.' Write 20-40 words.
25	4	A2	Writing about what you do	Write about where you work, what you do, if you like your job and why / why not. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words.
33	5	A2	Planning to attend a music festival	You want to write an email to your family to tell them about the music festival you’re going to. Tell them about the date, the cost of a ticket, the equipment you’re taking and the music you want to listen to. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words. Use future forms such as: ‘I’m going to go to a music festival’, ‘I’m taking a tent, T-shirt and shorts…’, ‘I’m going to listen to pop and dance music’ and ‘Maybe it’ll rain’.
41	6	A2	Writing a movie plot	Decide what happens to John and Isabella. Write the final part of the story for your friend. Type into the input box. When you're finished, click 'Submit.' Write 50-70 words.

Task	Level	CEFR	Topic	Prompt
49	7	B1	Giving instructions to play a game	You're working at a summer camp for children aged 8-12. You have been given instructions for three popular camp games, but they are a little too difficult for the kids to understand. Your task is to write simpler instructions for them. Type into the input box. When you're finished, click 'Submit.' Write 70-100 words.
57	8	B1	Writing a natural remedies pamphlet	Read grandma’s email and choose three of the remedies you think will sell well. Write a pamphlet explaining the benefits of each product, and how to use it. You should explain who the product is best for. Type into the input box. When you're finished, click 'Submit.' Write 70-100 words.
65	9	B1	Giving feedback to a restaurant	You've reached the end of the survey form and are being asked to give additional information. Write down in your own words what you thought about the food and drinks. Type into the input box. When you're finished, click 'Submit.' Write 70-100 words.
73	10	B2	Helping a friend find a job	Send Anna the zookeeper's job ad. It deals with animals, it's outside and it looks exciting! Write an email to Anna encouraging her to apply for the job. Try to use words and phrases such as 'absolutely', 'totally', 'by far the...', 'amazing', 'exhilarating', 'urge' and 'encourage'. Write 100-150 words. Begin your email like this: Hi, Anna! I’ve found an absolutely amazing job for you. Let me tell you why you should apply...
81	11	B2	Writing a movie review	There is movie festival in your local town. Think of the last movie you watched and enjoyed. Write a review of the movie to promote the movie in the local newspaper. Write 100-150 words. Type your answer into the input box.
89	12	B2	Turning down an invitation	Write the email to Graham, politely declining his invitation. Make sure that you use polite phrases, explain why you can’t come, and invite him and his wife next week instead (suggest some possible evenings). Write 100-150 words. Type your answer into the input box.

Task	Level	CEFR	Topic	Prompt
97	13	C1	Writing a campaign speech	Write your campaign speech for student council president. Think about the profile you fit – are you a socialist or a capitalist, or a bit of both? Convince your classmates that you are the best candidate. Write 150-180 words. Type your answer into the input box.
105	14	C1	Writing advertising copy	Choose one of the 3 images and slogans. Write an email to the CEO, giving reasons why you've selected the particular image and slogan and how it fits with Century's image of classic, stylish products. Write 150-180 words. Type your answer into the input box.
113	15	C1	Covering a news story	Task 1: You're a journalist covering a big murder trial. Your connection at the police station has given you the interview tapes. Listen and make notes about what happened. Task 2: It's time to write your report. Write the story, adding some details of your own to 'spice it up'. Use the audio to help you. Write 150-180 words. Type your answer into the input box.
121	16	C2	Attending a robotics conference	Write a short report comparing the three robots for the web-based magazine. Write 150-180 words. Type your answer into the input box.

Learners generally complete the tasks at home, with no time limit. Learners might consult their notes, though the large number of spelling and grammatical errors in the data indicate that they do not generally use tools such as spell-checkers. Furthermore, learners generally pay for EF's program themselves, and engage in the program for their own learning, so they are motivated to make the most of it for learning purposes (rather than cheat).

The texts that learners write are graded and given feedback by a teacher, and must receive a passing grade for the learner to advance to the next lesson.[51] The EFCAMDAT contains a mixture of longitudinal and pseudo-longitudinal data, since learners generally complete only parts of the program (e.g., because they were placed at a high initial proficiency level), and since some may have continued the program after the data-collection period for the dataset has ended. When it comes to metadata, the EFCAMDAT lists learners' English proficiency and nationality, and learners are only added to the database if their nationality

[51] Although, as noted in §7.2.3.1, the EFCAMDAT contains almost no texts that were written by the same learner for the same task, and those few cases do not necessarily appear to be a case of the learner resubmitting their task.

matches their country of residence (Alexopoulou et al., 2017). Accordingly, past research that used the EFCAMDAT relied on learners' nationality to estimate their L1, an approach that been validated empirically (Alexopoulou et al., 2017; Y. Huang et al., 2018; Malmasi & Dras, 2015; Murakami, 2014), and also used in associated studies on other language datasets (Rabinovich et al., 2018).

7.2.2 EFCAMDAT tasks and lexical choices

There is growing evidence that task effects can play an important role in L2 lexical choices (Kyle et al., 2016; Michel, 2017; Reid, 1986; Zenker & Kyle, 2021), including in the EFCAMDAT (Alexopoulou et al., 2017; Michel et al., 2019). In the context of the EFCAMDAT tasks, learners' lexical choices can be influenced by various factors, including:

- The *content* expected to be included in the task's writing (e.g., the topic that learners are expected to write about).
- The *style* expected of the task's writing (e.g., the level of formality expected of learners).
- The task's material (e.g., the vocabulary words that appear in the prompt, in example answers if learners choose to view them, and in any background content that learners are writing about).
- The material in the preceding lesson (e.g., vocabulary words that learners just learned).

Tasks may also differ in how constrained they are, in the sense that some allow for more spontaneous and open responses, whereas others necessitate a narrower range of responses. However, all tasks can generally be viewed as fairly constrained, in the sense that they are written in response to a specific prompt, and in an educational context where the key goal of communication is learning.[52]

The factors that can lead to task effects in the EFCAMDAT are common in many educational contexts. This also means that, even in contexts where not all of these factors apply, it is generally likely that at least some of them will apply. For example, even if students are asked to write an essay on whatever topic they choose (i.e., without a specific prompt), there are still likely to be both explicit and/or implicit expectations regarding which topics are appropriate for the school environment. This is also the case in many common non-educational contexts, such as the workplace, where similar factors can also lead to task effects.

[52] As noted above, learners generally pay for the course themselves, so it is in their best interest to use the opportunity in order to learn, rather than get a good grade.

There is limited publicly available information about many of these factors (e.g., about the vocabulary that learners were presented with in the lessons preceding each task), aside from tasks' topics and prompts, examples of which are presented in the previous sub-section. Accordingly, there is currently limited information regarding how these factors influence learners' productions in the EFCAMDAT. The most relevant of this information appears in Michel et al. (2019), who manually categorized tasks based on different task types (e.g., argumentative or descriptive), as well as in a similar preceding study by Alexopoulou et al. (2017).

In the present research, we include *task* as a random effect within our mixed-effects models, to control, in aggregate, for all aspects of each task that can influence lexical choices, such as prompt and the preceding lesson.[53] This approach does not attempt to disentangle the different aspects of each task that lead to task effects, so we will make no claim regarding the impact of any specific aspect of tasks on learners' lexical choices.

Finally, note that, despite these effects, past studies on the EFCAMDAT were able to find L1 effects in a wide range of linguistic phenomena. This includes lexical transfer (X. Jiang et al., 2014), capitalization (Shatz, 2019), articles (Shatz, 2017), grammatical morphemes (Murakami, 2016), clause subordination (Chen et al., 2021), relative clauses (Alexopoulou et al., 2015), and clause-initial prepositional phrases (X. Jiang et al., 2014).

7.2.3 The EFCAMDAT Cleaned Subcorpus

The present sample comes from the EFCAMDAT Cleaned Subcorpus,[54] which was derived from the full EFCAMDAT as outlined in Shatz (2020). The key feature of this dataset is that it is split into two sub-corpora, each containing texts written by similar learners in response to different prompts. This means, for example, that both the first and second sub-corpora contain texts written by German learners in task #4, but the learners in the first sub-corpus wrote their texts in response to a different task prompt than the learners in the second sub-corpus, and it occurred as result of a partial update of the Englishtown content during the data collection period. As such, using this dataset presents two important advantages for research. First, it allows us to more accurately categorize texts based on the task that they belong to, which leads to more accurate assessment of task effects. Second, as noted by Shatz (2020), this type of

[53] The only exception is the task's associated L2 proficiency level, which we control for using the relevant predictor, as explained in the paper.

[54] As accessed/downloaded on 4-January-2021.

dataset offers an opportunity to conduct our analyses on two similar but distinct learner samples, which serves as a form of replication.

The EFCAMDAT Cleaned Subcorpus contains texts from the A1–C1 CEFR range, and from the top 11 nationalities with most texts in the EFCAMDAT (referred to henceforth as “L1s”, rather than “nationalities”, since, as discussed earlier, we are using nationality to estimate L1). The distribution of texts for each combination of L1 and CEFR level is shown in Table 13.

Table 13. Number of texts in the EFCAMDAT Cleaned Subcorpus, per L1 and CEFR proficiency level. L1s are listed by the total number of texts that they have in the first corpus, in decreasing order. Data is taken from Shatz (2020), which outlines the creation process of this dataset. L1 is estimated here based on learners' nationality, as discussed earlier; in cases where the name of the L1 is different than the name of the nationality, notes have been added to clarify ambiguities.

L1	First corpus						Second corpus					
	Total	A1	A2	B1	B2	C1	Total	A1	A2	B1	B2	C1
Portuguese [a]	149,297	75,497	45,407	20,989	5,830	1,574	164,241	85,191	42,105	25,520	9,412	2,013
Mandarin [b]	86,660	45,008	29,318	10,321	1,763	250	20,317	10,494	6,021	2,730	936	136
Spanish [c]	34,559	19,296	9,847	4,102	1,114	200	30,204	15,998	7,645	4,500	1,740	321
Russian	32,243	12,295	10,885	6,329	2,066	668	17,078	7,249	4,652	3,449	1,443	285
German	24,705	8,041	7,860	5,051	2,698	1,055	16,717	4,652	4,487	4,083	2,669	826
French	19,135	7,626	6,253	3,688	1,242	326	13,384	4,610	3,755	3,188	1,528	303
Italian	18,959	5,899	6,832	4,291	1,466	471	16,469	5,046	5,010	4,166	1,749	498
Arabic [d]	13,152	7,463	3,729	1,412	417	131	16,156	8,089	4,874	2,301	727	165
Mandarin [b]	11,711	4,116	4,298	2,506	650	141	10,900	3,668	3,731	2,490	893	118
Japanese	9,149	3,337	3,095	1,903	640	174	7,937	2,812	2,409	1,837	701	178
Turkish	6,492	3,085	2,067	914	301	125	3,817	1,683	1,064	769	253	48
Total	*406,062*	*191,663*	*129,591*	*61,506*	*18,187*	*5,115*	*317,220*	*149,492*	*85,753*	*55,033*	*22,051*	*4,891*

[a] This is based on the Brazilian nationality (i.e., Brazilian Portuguese).

[b] The first entry of Mandarin, which contains more texts, refers to Chinese Mandarin (i.e., the Chinese nationality); the second entry refers to Taiwanese Mandarin (i.e., the Taiwanese nationality).

[c] This is based on the Mexican nationality (i.e., Mexican Spanish).

[d] This is based on the Saudi Arabian nationality.

7.2.3.1 Differences and similarities between the two corpora

As noted previously (in the section presenting background information on the EFCAMDAT), information is available on the prompts that learners wrote their texts in response to. However, this information is only available for texts in the first corpus in the EFCAMDAT Cleaned Subcorpus, since the separation of the original EFCAMDAT into two corpora was programmatic, as outlined by Shatz (2020), so listed prompts were available only for the first corpus, but not the second.

Nevertheless, this is not crucial to the present research, for several reasons. First, to the best of our knowledge, the texts were produced in the same general educational environment, with the same general learning goals, by similar learners. This is supported by the fact that the writings in the two corpora are largely similar, for example in terms of wordcount, as shown in Figure 14 and Table 14. Furthermore, the way we controlled for task effects as random effects in our models minimizes any issues associated with this, since we make no claim about any specific aspects of the tasks, but rather treat all their aspects in aggregate (as noted earlier and as discussed in more detail in the Methodology section of the paper). Finally—and most importantly—our results replicated with high similarity across the two corpora (especially when it comes to the main mixed-effects models), indicating that any differences between them did not change our key findings.[55]

[55] One difference between the two corpora that we do know about is that there are 8 tasks per EFCAMDAT level in the first corpus and 6 tasks per level in the second, as discussed in the supporting documents of Shatz (2020). But, like other potential differences in prompts, this did not generally appear to substantially affect the key findings across the two corpora.

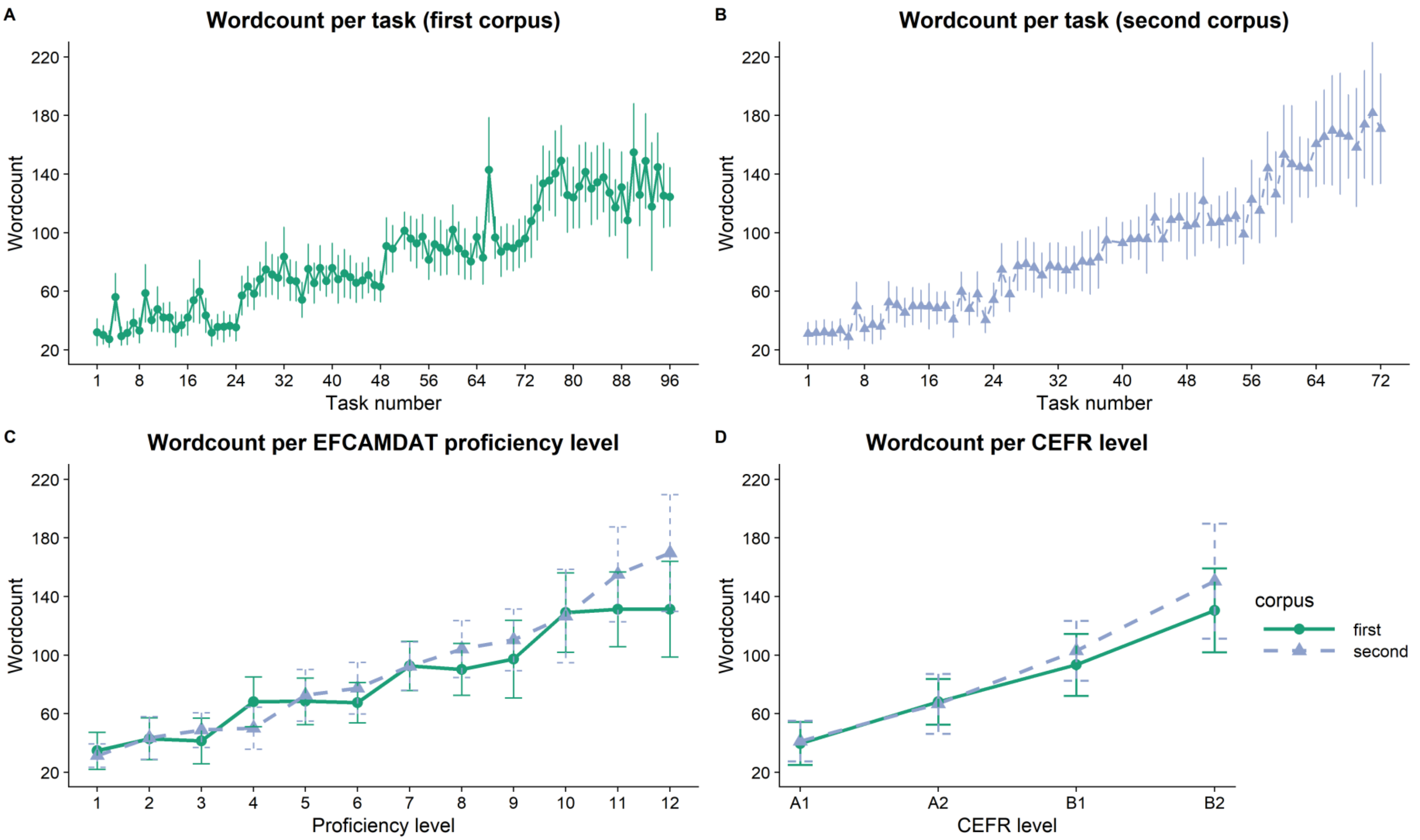


Figure 14. Mean wordcount of texts (error bars indicate one standard deviation), in the final samples that were selected for the study, as outlined in the next sub-section (N = 8,500 in the first corpus and 6,390 in the second). Listed per *task* in (A) and (B), per *EFCAMDAT proficiency level* in (C), and per *CEFR level* in (D). There are 8 tasks per EFCAMDAT proficiency level in the first corpus and 6 tasks per level in the second. There are 3 EFCAMDAT levels per CEFR level in both corpora.

Table 14. Statistics on the wordcounts of texts per CEFR (L2 proficiency) level in each corpus, in the final samples that were selected for the study, as outlined in the next sub-section (*N* = 8,500 in the first corpus and 6,390 in the second). *SD* denotes standard deviation, and *n* denotes the number of texts at that CEFR level.

	First corpus						Second corpus					
CEFR	Mean	SD	Median	IQR	Range	n	Mean	SD	Median	IQR	Range	n
A1	39.71	14.63	37	29-45	20-111	2160	41.25	13.87	40	30-50	20-105	1620
A2	68.11	15.59	68	58-75	22-133	2160	66.78	20.41	65	52-80	20-129	1620
B1	93.41	21.15	93	79-101	28-197	2065	102.97	20.50	104	91-115	31-188	1530
B2	130.60	28.56	130	109-149	21-245	2115	150.54	39.20	153	123-177	31-287	1620

7.2.4 Sample selection process and final sample

Two of the 11 nationalities that appear in the EFCAMDAT Cleaned Subcorpus data were excluded from the present study:[56]

- *Turkish*, since it had a too few texts for analysis, especially for certain tasks at the upper proficiency levels (see Table 13 for the number of texts per CEFR level written by Turkish speakers).
- *Taiwanese,* since both it and the *Chinese* nationality had data for the *Mandarin* L1, but the Chinese nationality had substantially more texts (as shown in Table 13). The two samples were not aggregated, because the Chinese nationality was already one of the nationalities with the most texts (2nd most in the first corpus and 3rd most in the second corpus), so adding the texts from the Taiwanese nationality would not have helped given the relatively balanced selection process of texts that is outlined below, and would have only caused a potential confound in terms of the learner backgrounds (Nisioi, 2015).

In addition, the C1 CEFR level (corresponding to EFCAMDAT levels 13-15) was excluded, since there were relatively few texts at this level, particularly for the L1s with the lower number of texts in general. Finally, texts were removed if they were not the first text that a learner submitted for a certain task.[57] This included only 1,045 texts in the first corpus (0.27%) and 284 texts in the second corpus (0.10%), after excluding the Turkish, Taiwanese, and C1 texts.

After this initial process, we selected a random subset of texts from the dataset, in a way that keeps the number of texts relatively balanced across L1s, proficiency levels, and tasks. This was necessary due to the extreme differences between the number of texts available for different L1s/proficiency levels/tasks, which were substantial enough to lead to issues with the interpretability of the findings.[58] In this regard, it is important to note that the initial differences in the distribution of texts is primarily reflective of EF’s market considerations, for example when it comes to the countries where they are most active. As such, selecting a relatively

[56] This selection of a subset of nationalities from the sample is in line with past research that used the EFCAMDAT, as in the case of Murakami (2016) and Shatz (2019), who examined only the top 10 nationalities with most texts, and Alexopoulou et al. (2015) and Geertzen et al. (2014), who examined only the top 5. This is also in line with the EFCAMDAT Cleaned Subcorpus itself, which, as noted here, contains only texts from the top 11 nationalities with most texts.

[57] It is not possible to know what led to some of these cases, such as the case where learner #174187 wrote 5 separate texts for task #80 over the course of almost a year, though the initial text had a grade of 87.

[58] For example, the Portuguese sample in the first corpus has ~75,500 texts at the A1 level, compared to only ~5,800 texts at the B2 level, and more crucially, compared to only ~3,300 texts at the Japanese A1 level, and only ~650 texts at the Japanese B2 level. Similarly, in the second corpus, Portuguese had more texts in total than all the other L1s combined.

balanced subset is a reasonable way to eliminate the substantial imbalance that appears in the original sample, and is in line with many past studies on the EFCAMDAT, who used similar subsetting procedures, and worked on specific and relatively balanced subsets of the corpus (Geertzen et al., 2014; Y. Huang et al., 2020; Malmasi & Dras, 2015; Murakami, 2016; Nisioi, 2015), as well with studies on other learner samples, such as the *TOEFL11* corpus (Kyle et al., 2015).

The final sample that was selected for the study is outlined in Table 15. It contains 10 texts for each combination of task and L1 (e.g., 10 texts written by German speakers in task #4), with a few exceptions, which are outlined in §7.2.5.1. There are 8 tasks per CEFR level in the first corpus and 6 tasks per CEFR level in the second corpus.[59] Accordingly, in the first corpus, there are 207–240 texts for each of the 9 L1s per each of the 4 CEFR levels, and 8,500 texts in total. In the second corpus, there is the same number of L1s and CEFR levels, and there are 170–180 texts per combination of L1/CEFR level, and 6,390 texts in total.

[59] The one exception is task #51 in CEFR B1, which was removed from both corpora in the sample, because texts from both corpora in this task were classified under the first corpus, due to limitations in the classification scheme that was used in the EFCAMDAT Cleaned Subcorpus (Shatz, 2020).

Table 15. Number of texts in the final sample, per each combination of L1/CEFR proficiency level.

L1	First corpus					Second corpus				
	Total	A1	A2	B1	B2	Total	A1	A2	B1	B2
Arabic	915	240	240	228	207	710	180	180	170	180
French	950	240	240	230	240	710	180	180	170	180
German	950	240	240	230	240	710	180	180	170	180
Italian	950	240	240	230	240	710	180	180	170	180
Japanese	939	240	240	227	232	710	180	180	170	180
Mandarin	949	240	240	230	239	710	180	180	170	180
Portuguese	950	240	240	230	240	710	180	180	170	180
Russian	950	240	240	230	240	710	180	180	170	180
Spanish	947	240	240	230	237	710	180	180	170	180
Total	*8,500*	*2,160*	*2,160*	*2,065*	*2,115*	*6,390*	*1,620*	*1,620*	*1,530*	*1,620*

Note. There are 8 tasks per CEFR level in the first corpus and 6 tasks per CEFR level in the second corpus. The one exception is CEFR level B1, where task #51 was removed both corpora in the present sample, because texts from both corpora in this task were classified under the first corpus due to limitations in the classification scheme that was used in the EFCAMDAT Cleaned Subcorpus (Shatz, 2020).

7.2.5 Additional sample information

7.2.5.1 Cases with fewer than 10 texts

Table 16 contains cases where there were fewer than 10 texts at a certain task for a certain L1 (e.g., at task #64 for Arabic). There were only such cases in the first corpus, and only 14 (1.64%) such cases out of 855 combinations of task and L1. All but one case had 5 or more texts (mean = 6.43, SD = 1.79, range = 2–9). Given the small number of these cases, and the small difference that they generally had in terms of the number of texts compared to the regular cases in the sample with 10 texts, this does not substantially influence the analyses or the interpretation of the findings.

Table 16. Cases with fewer than 10 texts at a certain task for a certain L1.

Task Number	L1	Number of texts
64	Arabic	8
64	Japanese	7
85	Arabic	8
86	Arabic	8
87	Arabic	5
88	Arabic	7
92	Arabic	6
93	Arabic	7
93	Mandarin	9
94	Arabic	6
95	Arabic	5
96	Arabic	5
96	Japanese	2
96	Spanish	7

7.2.5.2 Number of texts per learner

Table 17 contains the number of texts written per learner in the sample. In most cases, learners had only a single text in the corpus, though some learners had more than that. The main reason for including cases where a learner had more than one text is to ensure that we have a sufficient number of texts for analysis across all the combinations L1/task in the sample. This style of analysis is in line with previous studies on the EFCAMDAT, which also included multiple texts per learner (e.g. Alexopoulou et al., 2017; Michel et al., 2019; Shatz, 2019).

Table 17. Number of texts per learner. Note that in all cases, both the *median* and the and *minimal* texts per learner were 1.

	First corpus					Second corpus				
			texts per learner					texts per learner		
L1	texts	learners	mean	SD	max	texts	learners	mean	SD	max
Arabic	915	487	1.88	2.32	24	710	485	1.46	1.11	9
French	950	571	1.66	1.34	11	710	462	1.54	0.96	7
German	950	674	1.41	0.88	8	710	495	1.43	0.84	7
Italian	950	580	1.64	1.28	13	710	472	1.50	0.94	7
Japanese	939	448	2.10	2.15	17	710	382	1.86	1.50	11
Mandarin	949	755	1.26	0.70	7	710	524	1.35	0.95	9
Portuguese	950	810	1.17	0.54	8	710	665	1.07	0.26	3
Russian	950	696	1.36	0.81	7	710	516	1.38	0.93	9
Spanish	947	594	1.59	1.45	12	710	511	1.39	0.85	8
Overall	*8500*	*5615*	*1.51*	*1.33*	*24*	*6390*	*4512*	*1.42*	*0.96*	*11*

7.3 Appendix C: Lexical distance & baseline frequency information

This document contains information regarding the lexical-distance datasets that were used in the study, as well as regarding how the baseline frequency of English words was calculated. The majority of the content here applies to study 2 (on word choice), though the material about the Swadesh lists also applies to study 1 (on lexical diversity).

7.3.1 Brief overview of the lexical-distance datasets

7.3.1.1 Swadesh lists

Swadesh lists are lists of words that represent concepts that appear in nearly all languages, such as *water*, *night*, *full*, and *hear*, with each list containing the equivalent of those words in a specific language. These lists are often used in lexicostatistical studies, primarily to calculate the distance between languages (Bakker et al., 2009; Holman et al., 2008a; Schepens, van der Slik, et al., 2013b; Swadesh, 1950, 1955; Wichmann et al., 2010, 2011).

Table 18 contains a sample from several Swadesh lists, taken from the *Automated Similarity Judgment Program* (ASJP, version 18) (Wichmann et al., 2018).[60] This resource contains a large selection of Swadesh lists for various languages, which is often used for studies on the topic (Bakker et al., 2009; Holman et al., 2008a; Schepens, van der Slik, et al., 2013b; Wichmann et al., 2010, 2011).[61]

[60] New versions of the ASJP are released periodically, with added data and other modifications. The present study used data from version 18 of the dataset, though version 19 has been released since then.

[61] The frequent use of this source also means that it has been extensively validated. This includes studies that cross-validated Levenshtein distances that were calculated from these lists, by comparing them with other measures of language distance, and showing that they strongly correlate with expert-based cognancy judgments (Schepens, van der Slik, et al., 2013b); psychoacoustic, psycholinguistic, phonetic, and phonological measures of distance (Brown et al., 2013); taxonomic distances in the *World Atlas of Language Structures* (WALS) and the *Ethnologue* (Holman et al., 2008b); and distances based on morphological and typological features in the WALS (Bakker et al., 2009; Schepens, van der Slik, et al., 2013a). In addition, further support for these lists comes from studies that used it for language classification and linguistic phylogenetics, and involved various forms of validity and reliability checks (Brown et al., 2008; Holman et al., 2008b, 2008a; Pompei et al., 2011; Wichmann, 2019; Wichmann et al., 2010). Finally, distances from these lists were also used successfully in the context of SLA, to predict Dutch L2 proficiency (Schepens, van der Slik, et al., 2013b, 2013a), which aligns with studies that predicted L2 Dutch proficiency using two different types of cognancy judgments (Schepens et al., 2020; van der Slik, 2010).

Table 18. A selection of words from Swadesh lists in English, German, French, and Japanese.

	Phonetic transcription			
Meaning	English	German	French	Japanese
I	Ei	ix	j3	wataSi
you	yu	du	ti	anata
one	w3n	ains	oe*	hitocu
fish	fiS	fiS	pw~aso*	uo
blood	bl3d	blut	sa*	Ci
horn	horn	horn	korn	cuno
ear	ir	or	ore	mimi
drink	drink	triNk3n	bw~a	nomu / su
come	k3m	kh~om3n	v3ni	ku / yuku
sun	s3n	zon3	sole	suna
full	ful	fol	pl3*	miCita / ippai de
new	nu	noi	nuvo	ataraSi / Sinsen

Note. The ASJP uses a specialized set of characters in its transcription, as outlined in Brown et al. (2008). One important limitation of this transcription is that it collapses certain phonological distinctions, since in some cases, different underlying segments are transcribed using the same character; for example, the "voiced bilabial stop and fricative" (IPA: /b/ and /β/) are both transcribed using /b/ in the ASJP. In addition, as with similar source, there may be issues with the ASJP transcriptions, since some words may be transcribed incorrectly, and sometimes the wrong words be may selected for transcription. Accordingly, despite the extensive validation for these lists that was noted above, and as with distances based on similar sources, distances that are based on the ASJP should preferably be interpreted with caution, be used for large-scale analyses that can accommodate some noise, and be connected to follow-up analyses that use alternative sources of distance, as we do in the present study.

The Swadesh lists in the ASJP contain up to 100 standard concepts (i.e., general meanings) per language. However, the ASJP is focused on a subset of 40 concepts out of the initial 100, and "As a rule of thumb the database normally only includes lists that are at least 70% complete, i.e., which contain at least 28 items on the 40-item list" (Wichmann et al., 2011, p. 3). This is based on studies that showed that this subset represents the most stable elements from the original list, whose use leads to optimal results when it comes to language classification (Bakker et al., 2009; Holman et al., 2008b, 2008a). The following list represents all the concepts in the database, with the subset of 40 main concepts in bold:

> **I**, **you**, **we**, this, that, who, what, not, all, many, **one**, **two**, big, long, small, woman, man, **person**, **fish**, bird, **dog**, **louse**, **tree**, seed, **leaf**, root, bark, **skin**, flesh, **blood**, **bone**, grease, egg, **horn**, tail, feather, hair, head, **ear**, **eye**, **nose**, mouth, **tooth**, **tongue**, claw, foot, **knee**, **hand**, belly, neck, **breast**, heart, **liver**, **drink**, eat, bite, **see**, **hear**, know, sleep, **die**, kill, swim, fly, walk, **come**, lie, sit, stand, give, say, **sun**, moon, **star**, **water**, rain, **stone**, sand, earth, cloud, smoke, **fire**, ash, burn, **path**, **mountain**, red, green, yellow, white, black, **night**, hot, cold, **full**, **new**, good, round, dry, **name**
>
> From "ASJP" (2018)

7.3.1.2 Parallel dictionaries

The parallel dictionaries are similar to the Swadesh lists, in the sense that they are lists of words in different languages that can be directly compared between different languages. The parallel dictionaries in this case came from the *Intercontinental Dictionary Series* (IDS) (Key & Comrie, 2015).[62] The entries in the IDS are organized in 22 chapters, each of which revolves around a topic such as animals (e.g. 'bird'), the body (e.g. 'head'), the house (e.g. 'wall'), time (e.g. 'immediately'), cognition (e.g. 'learn'), and law (e.g. 'acquit'). The IDS was modelled after *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*, which was compiled by Carl Darling Buck, which contained approximately up to 1200 general entries per language. The IDS itself contains 1310 general entries (i.e. general word meanings, not taking the number of synonyms into account), and if a certain form does not exist in a certain language, the entry for it is left blank (Key & Comrie, 2015). It offers less diversity in terms of the number of languages that includes, compared to the Swadesh lists, but far more words per language.

7.3.2 Processing the lexical-distance datasets

7.3.2.1 Swadesh lists

The words in the Swadesh lists in the *Automated Judgment Similarity Program* (ASJP) are transcribed using a specialized phonetic script (outlined in Brown et al., 2008), so these transcriptions were converted to IPA using the dedicated *asjp* library in Python (Sofroniev, 2018).

[62] No version number could be found for the database; data was downloaded from it on 04-Dec-19.

In addition, due to the focus on single-word entries here, we removed 15 entries (4.39% of 342) that contained multi-word phrases, either in the English meaning or in the corresponding L1 translations.[63] In addition, to control for the number of meanings across L1s, as was done during the initial preparation of the lexical-distance data, single-word entries were removed if they contained the same English meaning as an entry in another L1 that had a multi-word meaning. For example, if the French entry for a certain English meaning was removed for being multi-word, then the entries in the other L1s that correspond to the same meaning were also removed. Based on this, a further 102 entries were removed (31.19% of 327). As such, the final version of the Swadesh list used for analysis contained 225 entries, with 25 entries for each of the 9 L1s, where each entry is a row containing an English word an all its corresponding L1 translations.

7.3.2.1.1 Arabic dialect

According to Eberhard et al. (2020), the three most common Arabic dialects in Saudi Arabia are:[64]

- Spoken Najdi Arabic (the *de facto* national working language, with ~14,600,000 speakers out of a population of ~33,414,000).
- Spoken Hijazi Arabic (~10,300,000 speakers).
- Spoken Gulf Arabic (~962,000 speakers).

In the version of the ASJP that was used (18), there was data for two of these dialects: Najdi Arabic and Gulf Arabic.

However, Najdi Arabic had substantially fewer entries available than the other L1s (30 general word meanings, compared to 40–100). Because we wanted to use the same words across all L1s, to minimize potential confounds as a result of using different words, we decided

[63] One issue with analyzing these phrases is that there are many cases where the same word appears both by itself and as part of a multi-word phrase (e.g. *hand* and *palm of hand* in the parallel dictionaries). Another issue is that the frequency data for multi-word phrases is estimated, unlike the frequency data for single-word phrases, which is calculated directly based on existing datasets. Furthermore, most of the multi-word phrases are expected to be very rare in writing, to a point where it will be difficult to assess their acquisition. Finally, from a linguistic perspective, it is expected that these phrases will have different acquisition patterns than the single-word entries, as they are more likely, for example, to be influenced by syntactic patterns that relate to how the words are used together. While it is possible to account for some of these issues, and while there is value in studying the acquisition of multi-word phrases and in comparing it to the acquisition of single words, we decided to focus on the acquisition of single words in the present study, in line with most prior research.

[64] There is also Standard Arabic, which is the national language in Saudi Arabic, but according to Eberhard et al. (2020), it is “Not an L1. In most Arab countries only the well-educated have adequate proficiency in Modern Standard Arabic.”.

to use Gulf Arabic instead of Najdi Arabic in the present sample, since using Najdi Arabic would have reduced the number of words for each L1 in the sample.

An analysis showed that there is substantial similarity between the entries that the two dialects of Arabic share in the Swadesh lists, as the correlation between the LDN from the closest synonym in each dialect was *Pearson's r* = .5438, *p* = .0011. Furthermore, the mean LDN for Najdi Arabic was 0.90 (SD = 0.13, median = 1) and for Gulf Arabic was 0.92 (SD = 0.12, median = 1), which, with less rounding, is equivalent to only 0.0182 (|0.8978−0.9160|).

In practice, if Najdi Arabic was used as-is, without modifying the words in the other L1s, then Arabic would have only shifted from being ranked the 8th most distant L1 to the 7th most distant L1, by switching places with Japanese (mean LDN = .91), which is a relatively small difference. Furthermore, given that Arabic was ranked #9 and #8 for mean MTLD (in the first and second corpora), this change would have supported (and slightly strengthened) the current conclusions of the study, regarding the lack of association between lexical distance and lexical diversity, since there would have been a bigger gap between Arabic's rank in terms of lexical distance and its rank in terms of MTLD.

Overall, while it would have been ideal to be able to use Najdi Arabic, these analyses suggest that Gulf Arabic was a reasonable substitute in the present study.[65]

7.3.2.2 Parallel dictionaries

The second lexical-distance dataset is the *Intercontinental Dictionary Series* (IDS), which contains parallel dictionaries in various languages—essentially, lists of words of corresponding words in various languages, similarly to the Swadesh lists (Key & Comrie, 2015).[66] The parallel dictionaries required more processing before they could be used in the analyses, as shown in the following sub-sections. In addition, the parallel dictionaries contain data for fewer languages than the Swadesh lists, but for more words per language, as shown in the following sub-sections, so it is beneficial to use this dataset in the present study, to complement the Swadesh lists.

[65] A suggestion to consider for future research is to instead use the data from Najdi Arabic (rather than Gulf Arabic), and only substitute the missing values using Gulf Arabic (or preferably, using Hijazi Arabic, as it was added to a later version of the ASJP).

[66] No version number could be found for the database; data was downloaded from it on 04-Dec-19.

7.3.2.2.1 Initial cleanup process

Data was initially available in the IDS for the following languages (out of the ones that were included based on the learner sample): French, German, Italian, Portuguese, Russian, Spanish, and English. For all languages, entries were transcribed orthographically, with two exceptions. First, entries in Spanish were transcribed orthographically with a minor modification, which is explained and addressed later (§7.3.2.2.2). Second, entries in Russian were transcribed phonemically using a specialized non-IPA script, which could not be reliably transformed into IPA, so this L1 was excluded from this dataset.

Because this is a large-scale resource, with +1,500 entries per language in the sample, collected by different researchers at different periods of time, a substantial initial organization and cleanup process was required.[67] The goals of this process were to resolve programmatic issues that interfered with analyses (e.g., the removal of 'null' entries), to remove extraneous data which was artificially included as part of transcriptions and which could interfere with calculations of lexical distance (e.g., question marks), and to resolve other types of artificial issues that could interfere with the analyses (e.g., parenthetical information).[68] In doing so, the overarching goal was to reduce artificial noise and errors in the data, even at the cost of removing some data which was problematic in a way that could not be properly resolved. Overall, as shown below, only a relatively minor portion of the entries had issues that necessitated the removal of the full entry, and given this, together with the relatively large size of the dataset, this left sufficient entries for the main analyses.

The cleanup process included the following.

First, entries corresponding to the meaning of 'zero, nothing' (IDS code 13.0) were removed, because the German entry for 'zero' ('null') caused some software issues during analysis. Corresponding entries were removed from all language, rather than just from German, to maintain a balanced sample. This included a total of 2 entries in English ('zero' and 'null', which account for 0.13% of the 1,551 entries), and 11 entries across the other languages (0.13% of 8,507).

Then, question marks (? and ¿) were removed from the transcriptions, and the entries that contained them were marked as questions. In addition, in the French entry for 'which' (ID

[67] Initially, there were 1,551 entries in English and 8,507 entries in the 6 other languages, with a mean of 1701.40 entries in each language (*SD* = 135.09, range = 1539–1898).

[68] The term "artificial" here is used to denote the fact that the issues are attributed to this particular dataset and to the way it was collected, rather than to an underlying linguistic phenomenon.

17-670-171-1), which contained two separate words (“quel? lequel?”) within the same entry, the second word was removed.

Then, the entries for the meaning of *he/she/it* (IDS code 2.930), were removed, due to the inconsistent way they were formatted across different languages. For example, in French, the words were all listed under the same entry, similarly to English, while in Italian each word was listed as a different entry, under the same single meaning.[69] This included a total of 1 entry in English (0.06% of 1,549), and 12 entries in the all the other languages combined (0.14% of 8,496).

Then, entries containing a slash (/) were removed, since slashes were used inconsistently across the different entries (no remaining entries in English, 25 entries in the other languages, 0.29% of 8,484).[70] This included, for example, cases where the masculine/feminine versions of a word were listed under the same entry, in various formats (e.g. *primo/ prima* ‘cousin’ in Portuguese and *ciervo/a* in Spanish), despite the fact that the two versions were not necessarily represented in the other languages. In addition, this also included other uses of slashes, in various forms (e.g. *llevar a la espalda/ en el hombre* for ‘carry-on-shoulder’ in Spanish).

Then, entries containing parentheses were modified as follows:

- In 34 cases where an entry contained a single word or expression fully surrounded in parentheses, the parentheses were removed. All such cases were in German, and always for synonyms of an initial entry, suggesting that this scheme may potentially have been originally used to mark synonyms (e.g. *(Stier)*, a second entry for the meaning ‘ox’ in German).
- A few entries contained an addition in parentheses that appeared to be a prefix (3, all in German, e.g. *(er)wachen*, ‘wake up’) or a suffix (3, all in Spanish, e.g., *baɲar(se)*, a partly phonemic transcription for ‘bathe’). In all such cases, the parentheses were removed from the transcription, but the lexical material inside the parentheses was preserved, since such material was generally necessary to preserve the original meaning of the word.

[69] Furthermore, because in Italian there were two possible words listed for each of the original three meanings, it would not have been possible to modify the entries in a way that would fit under the existing classification scheme.
[70] Note that the reason why the *he/she/it* entries had to be removed separately is that they did not always include slashes.

- Finally, there were 19 entries where at least one word appeared inside of parentheses and at least one word appeared outside of them (all in non-English entries). Because there was significant variability in terms of what kind of material was included in parentheses in each case, these entries (0.22% of the 8,459 non-English entries at this stage) were removed.

In addition, there were 4 entries (in German) were there were square brackets around the entire word (e.g. *[Sommerküche]* ‘cookhouse’); in all cases, the square brackets were removed from the transcription.

In entries where there was one hyphen (-) or more, the hyphens were replaced with spaces. Hyphens were used in some cases in place of a space to separate words, but they did not appear to convey a meaning beyond that of a regular space, and their use was inconsistent, and did not follow any discernible criteria, both within each language, and across languages. For example, in English, ‘grass-skirt’ and ‘garden-house’, contained a hyphen, but ‘spider web’, ‘body hair’, ‘fish trap’, and ‘walking stick’ did not. Furthermore, while hyphens were sometimes used to separate words in French and Portuguese, they were not used for this purpose in any of the other L1s in the sample. Accordingly, hyphens were removed from 22 entries in English and 46 non-English entries (26 in French and 20 in Portuguese).

Some of the entries in the IDS contained various types of parenthetical information in the ‘meaning’ column (which was the English value for the entries). This included, for example, ‘calm (of sea)’, ‘lightning (as striking)’, ‘rain (noun)’, ‘burn (vb trans)’, ‘man (vs. woman)’, ‘young man (adolescent)’, ‘son-in-law (of a woman)’, ‘you (singular)’, ‘we (inclusive)’, ‘male (adj)’, ‘cattle (bovine)’, ‘tree (cf 08.600)’, ‘sow (2)’, ‘fin (dorsal)’, ‘nit (louse egg)’, ‘beget (of father)’, ‘smoke (tobacco)’, ‘mold (clay etc)’, ‘dear (costly, expensive)’, ‘share (distribute)’, ‘thick (in dimension)’, ‘little (quantity)’, few’, ‘long-time (for a)’, ‘think (= reflect)’, ‘think (= be of the opinion)’, ‘ask (question, inquire)’, ‘light (in color)’, ‘people (populace)’. Because of the wide range of reasons why this parenthetical information was included, and the range of ways in which it was, entries containing parenthetical information in their original meaning were removed entirely from the sample.[71] This included

[71] This is because there was no reliable way to handle the parenthetical information, and it was therefore preferable to remove it in order to reduce the amount of noise and issues in the dataset, even at the potential cost of the loss of some information, particularly because the parallel dictionaries contain sufficient relevant information even after such removal.

130 entries in English (8.40% of the 1,548 English entries at this stage), and a corresponding 727 entries in other languages (8.61% of the 8,440 non-English entries at this stage).

Finally, a small number of entries contained an apostrophe (') connecting two words for various reasons. This included 7 entries in English (0.49% of 1,418 entries), that has a possessive /s/ (e.g., *mother's brother*, *men's house*, and *one's native country*), and 23 entries in non-English languages (0.30% of 7,713 entries), and specifically in French (e.g., *chute d'eau* ‘waterfall’, *s'eveiller* ‘wake up’, and *lobe de l'oreille* ‘earlobe’) and Italian (5 e.g., *lobo dell'orecchio* ‘earlobe’ and *acino d'uva* ‘grape’). Because of the variability in the way these apostrophes and the related lexical material should be interpreted phonologically and morphologically, these entries were removed from the sample.

7.3.2.2.2 Dealing with non-standard orthography

The entries for most of the languages in the present sample were transcribed using standard orthography for each language. For example, the entries for the word ‘world’ were transcribed as *world* in English, *monde* in French, *mondo* in Italian, *mundo* in Portuguese, and *Welt* in German. However, there was one exception to this—Spanish—whose entries had to be modified accordingly.

Specifically, the transcription appeared to be in standard orthography for the most part. For example, the entry for ‘world’ was *mundo,* the entry for ‘charcoal’ was *carbón,* and one entry for ‘extinguish’ was *extinguir*. Nevertheless, to identify any possible variations, we used the following approach.

First, we calculated the frequency of the Spanish entries (i.e. how often they appear in Spanish), using the *wordfreq* library in Python (Speer et al., 2018); this package was also used to get the general frequency data for the entries, and will therefore be explained in more detail later. Then, we examined all the entries that had a frequency of ‘0’, which meant that they did not appear once in any of the extensive datasets used by *wordfreq*, since this suggests that the word is either extremely rare, or that it is spelled in an unconventional way (i.e., phonemically) in the present dataset.

A total of 43 entries (2.69% of 1,596 entries in Spanish) fit this criterion. An examination of them revealed that the only exception to standard orthography that could be identified was the use of /ɲ/ in place of /ñ/ (e.g., in *montaɲa* ‘mountain’). In all 38 entries where this occurred, the /ɲ/ in the orthographic transcription was replaced by an /ñ/.[72]

7.3.2.2.3 Dealing with capitalization

[Note: This is not relevant for the studies in the thesis, since they only looked at phonological distance. It was part of the initial cleanup and organization process that I conducted when working with the IDS, to facilitate future research on the resulting dataset, which may take orthographical distance into account.]

Some entries were capitalized in the original dataset for various reasons. This included the following:

- The entry for the meaning of *God* was capitalized for all L1s.
- A single Italian entry (*fico del Banian* ‘banyan’).
- In English, the entries for the word ‘I’ and for the days of the week (e.g., ‘Sunday’) were capitalized.
- In German, the entries for nouns were capitalized (e.g., *Welt* ‘world’), as is conventional in German.

Capitalization can influence calculations of orthography-based lexical distance, in cases where an upper-case letter is substituted with a lower-case one (and vice versa). Given the rarity of capitalization in the present dataset, as well as the fact that in one case all languages share the capitalization, and in another case the capitalization is in the target L2, this is unlikely to have a substantial influence on analyses, with the potential exception of German, where it is both relatively common and systematic. There, the presence of capitalization causes the orthographic lexical distance to be overestimated for German words, under the assumption that the substitution of an upper-case letter with its lower-case equivalent is less substantial than its substitution with a different letter. Conversely, removing capitalization causes the distance to be underestimated, by artificially eliminating some of the distances between the words.[73]

[72] Another solution that was considered for identifying these entries is a spellchecker such as the *pyspellchecker* library in Python (Barrus, 2019). However, this was much more likely to lead to false positives, likely because the spellchecker relies on the frequency of words to identify potential corrections, so it tended to “correct” rare words that are spelled correctly by changing them to a more common word that is spelled similarly.

[73] Another option is to used weighted Levenshtein distance, by giving substitutions of upper-case letters and their lower-case equivalents a lower cost than other substitutions (e.g., 0.5 instead of 1). This has the advantage of reflecting the unique nature of such substitution. However, this also introduces complexity and arbitrariness into the model, and makes it more difficult to interpret the findings in light of other findings in the field, given the relative rarity of this approach.

Accordingly, in the present study, we create a decapitalized version of the transcription, which may be used together with or instead of the original version.[74]

7.3.2.2.4 Final cleanup and organization

Once the cleanup of the entries themselves was completed, it was necessary to further clean up the remaining dataset, in order to minimize the influence of the prior removal of various entries from the sample, and ensure that the sample is balanced across all the L1s. Before explaining how this was done, it is beneficial to illustrate how entries in the dataset are structured. To start with a simplified example of how each entry looks, see Table 19.

Table 19. A simplified example of what entries in the dataset look like, intended to help illustrate their structure as it relates to the final stages of the dataset cleanup process.

Entry ID	Language	General Word ID	Synonym number	Entry	Meaning
1-215-170-1	Italian	1-215	1	sabbia	sand
1-215-170-3	Italian	1-215	3*	rena	sand
1-240-170-1	Italian	1-240	1	valle	valley
1-240-171-1	French	1-240	1	vallée	valley
4-870-171-1	French	4-870	1	médecin	doctor

Note. In the original database, synonym numbers are always odd, so that a number of ‘3’ indicates that the synonym is the second one for the general word ID, while ‘5’ indicates that it is third, and so on. ‘1’ indicates that this is the first entry for that general word ID.

Based on this, we see that an entry’s ID contains four pieces of information, based on the four numbers separated by hyphens:

- The first number marks the chapter to which the word belongs. This includes, for example, chapter 1, which is 'the physical world’.
- The second number marks the general underlying meaning that the entry corresponds to within a particular chapter. For example, meaning #215 in chapter 1 is ‘sand’. Combining this number together with the chapter number therefore gives the ‘general

[74] The one exception was to remove the capitalization that only appeared on the single Italian entry (*fico del Banian*) for consistency and because it appeared to be in error.

word ID', which is the general meaning that the entry corresponds to. Note that these numbers do not reflect the number of words in each chapter; for example, meaning #223 in a chapter might be followed by meaning #230, and then #240.

- The third number marks the entry's language. For example, language #170 is French.
- The fourth and final number marks which number synonym that the entry is, for that general word ID under that particular language. This number is marked by consecutive odds numbers starting from '1'. For example, in the case of the entries for the meaning of 'sand' in French, the first entry has the synonym number '1', while the second has the synonym number '3'.[75]

For example, if we look at entry with the id of '1-215-170-3', this means the following:

- The entry is in chapter '1' ('the physical world').
- The entry's main meaning under chapter 1 is '215' (so the general word ID is 1-215), meaning that it corresponds to 'sand' as a main meaning.
- The entry's language code is '170', meaning that it's in French.
- The entry's synonym number is '3', meaning that it's the second synonym for that underlying meaning in this particular language.

Based on this, the following types of entries were removed from the dataset, in the following order, as later summarized in Figure 15:[76]

- **Entries for which one of the synonyms was removed at an earlier stage of the analysis.** This means, for example, that if one of two synonyms for a certain general meaning was removed from the English dataset at an earlier point, then the remaining synonyms were removed at this stage. This was done on a per language basis, and was

[75] The original dataset contains two additional variables, that are not mentioned here, but that are sometimes used when working with the entries programmatically. These variables are: *pk*, which is a unique number that identifies each entry in the dataset, and *valueset_pk*, which is a unique number that identifies each set of synonyms in a language (so all words that are in the same language and correspond to the same general meaning have the same valueset_pk). For example, in Italian, entry 1-210-170-1 has pk of 32533 and valueset_pk of 26436, while entries 1-212-170-1 and 1-212-170-3 have pk's of 32534 and 32535 respectively, and a valueset_pk of 26437 for both. Essentially, the pk variable corresponds to the full ID of each entry, while the valueset_pk variable corresponds to the full ID minus the final number (the synonym identifier). Though pk and valueset_pk are less informative than the ID variable in its various forms, they can be beneficial in situations where one wants to work with a continuous integer variable rather than with a categorical one.

[76] In terms of order, the crucial thing is to remove entries for which one of the synonyms was removed at an earlier stage of the analysis before removing entries where the general meaning does not appear for all L1s, in order to ensure that all the L1s contain data for all the general meanings that are kept in the sample. For example, consider a situation where French and Italian both have entries for general meaning A, but French originally had two synonyms for it, one of which was removed, while Italian has only a single entry for it. If the cleanup isn't performed in the appropriate order, French will end up without this meaning, while Italian will keep it.

achieved by comparing the current version of the dataset with the original one, to identify all the entries where words with the same general meaning in that language were previously removed. In English, a total of 169 out of the original 1,551 entries (10.90%) were removed during the initial cleanup process, leaving 1,382 entries. Of these, a further 7 entries (0.51%) were removed at this stage, leaving 1,375 English entries. In the L1s, a total of 817 entries out of the original 8,507 entries (9.60%) were removed during the initial cleanup process, leaving 7,690 entries. Of these, a further 17 entries (0.22%) were removed at this stage, leaving 7,673 entries in the various L1s.

- **Entries where the general meaning does not appear for all L1s**. This issue too could occur as a result of prior removal of entries from the dataset. This means, for example, that if only half of the L1s contained an entry for a certain underlying meaning, that entry was removed. However, if all the L1s had entries for that general meaning but different numbers of synonyms, the entries were kept. In total, there were 1,191 unique general meanings in the dataset of the various L1s. Of these, 41 (3.44%) general meanings were not shared across all L1s. After accounting for these general meanings, 205 entries out of 7,673 (2.67%) were removed from the sample, leaving 7,468 entries.[77]
- **L1 entries that did not have a corresponding English translation.** This could happen, for example, due to a prior removal of an English entry from the dataset. 222 (2.97%) entries of the current 7,468 were removed as a result of this, leaving 7,246 entries. These entries represent 85.18% of the original 8,507 entries in the dataset.

[77] Note that multiple entries can be removed from a single general meaning under a single language, if several synonyms are listed for it.

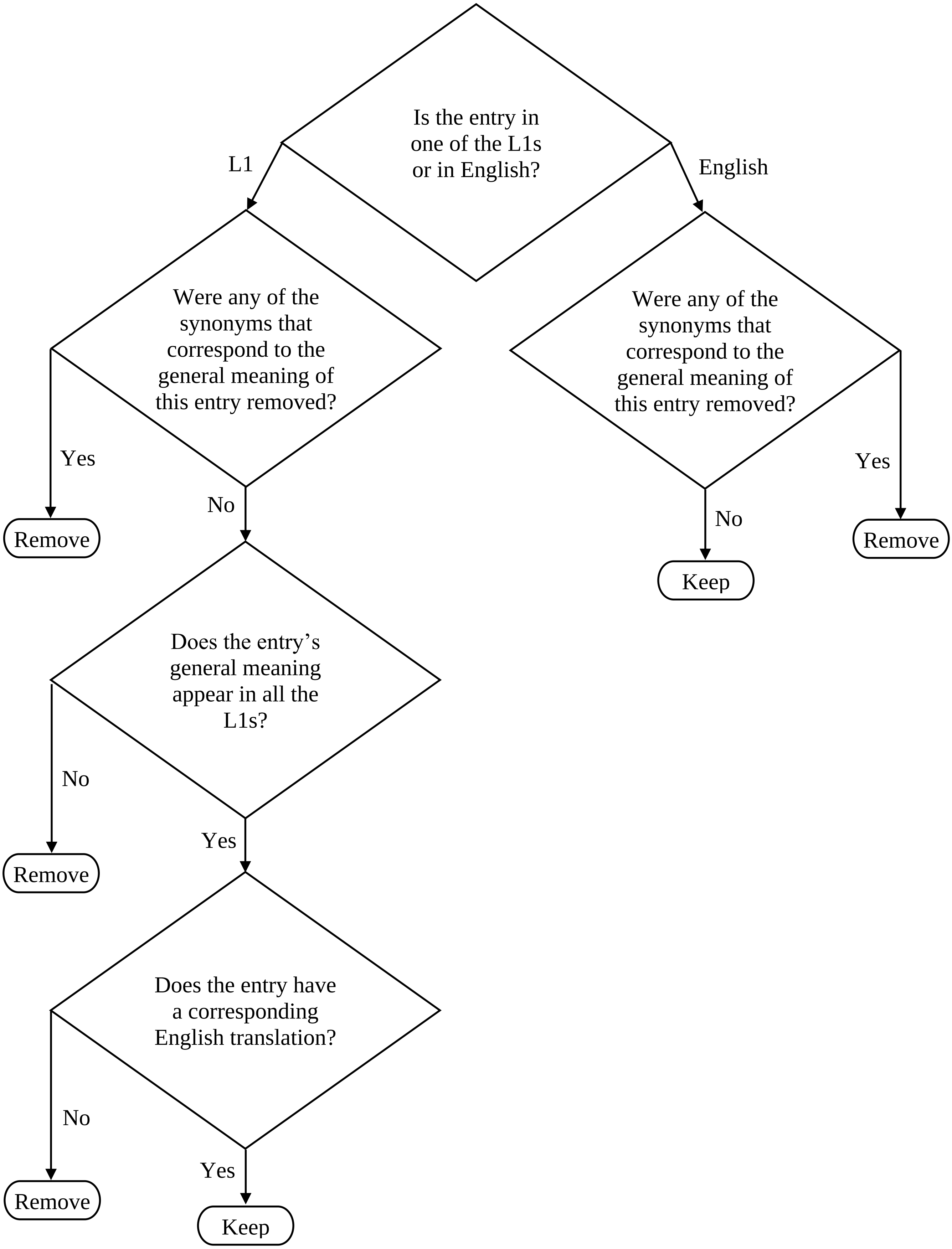

Is the entry in one of the L1s or in English?
L1
English
Were any of the synonyms that correspond to the general meaning of this entry removed?
Yes
No
Remove
Were any of the synonyms that correspond to the general meaning of this entry removed?
Yes
No
Keep
Remove
Does the entry's general meaning appear in all the L1s?
No
Yes
Remove
Does the entry have a corresponding English translation?
No
Yes
Remove
Keep

Figure 15. Flowchart showing the process used to determine which entries in the dataset will be removed, following the initial cleanup process.

Overall, the final 7,246 L1 entries in the dataset were divided between the five L1s. Each L1 had a different number of entries, as a result of a different number of synonyms, but all these entries belonged to an equal number of general word meanings (1,116). On average, there were 1449.20 entries for each L1 (*SD* = 118.37, median = 1,428, range = 1301–1619).

Note that the above information is based on when the entries were arranged based on L1, meaning that every row in the dataset contained the data corresponded to a single L1 entry, together with all its available English translations (with a separate row for each L1 synonym). When the entries were arranged to be English-based, meaning that each row corresponded to a single English entry together with all its corresponding entries in a single L1 (with a separate row for each English synonym), there were initially 6,695 entries in the table. This was after the first two stages of the cleanup process, where entries were removed if one of their synonyms was removed at an earlier stage of the analysis or if the general meaning of the entry did not appear for all L1s. Next, the final stage of the cleanup process was replicated, and English entries with no corresponding L1 translation were removed (45 entries, 0.67% of total), leaving 6,650 English entries, corresponding to 1,330 English meanings and 5 L1s. Overall, the total number of combinations of L1 and English entries, while taking synonyms into account, was 9,094.[78]

Since the present study focused on single-words entries, 514 multi-word entries (7.73% of 6,650) were removed from the sample, to avoid potential confounds, as with the Swadesh lists. Furthermore, to control for the number of meanings across L1s, single-word entries were removed if they contained the same English meaning as an entry in another L1 that had a multi-word meaning. Based on this, a further 401 entries were removed (6.54% of 6,136).

Next, in some cases, two entries in the parallel dictionaries (but not in the Swadesh lists) contained an English word that is spelled in the same manner (e.g., *drink* as a noun and a verb, *plain* as a noun and an adjective, *horn* as a body part and an instrument). Such entries were removed, because of the issues that they present when it comes to calculating how often the target word was used in the learners' writing. This included only 110 entries (1.92% of 5,735, which includes 22 meanings for each of the 5 L1s).

Finally, there were 22 cases where an entry appeared both in the Swadesh lists and in the parallel dictionaries. To ensure that the lexical material was unique to each dataset, and

[78] This is equivalent to the number of all English synonyms listed in the L1-based table, where each row represents an L1 entry, and to the number of all L1 synonyms listed in the English-based table, where each row represents an English entry.

since the Swadesh lists had fewer words than the parallel dictionaries, such cases were removed from the parallel dictionaries. This also included only 110 entries (1.96% of 5,625, which includes 22 meanings for each of the 5 L1s).

After this, there were 5,515 entries, with 1,103 entries for each of the 5 L1s, where each entry is a row containing an English word an all its corresponding L1 translations.

7.3.2.2.5 Generating IPA transcriptions

The parallel dictionaries' orthographic transcriptions of entries in each L1 were converted to IPA using Python's *epitran* library (version 1.8), which is a grapheme-to-phoneme (G2P) transduction system, dedicated to transliterating orthographic text in various languages into IPA (Mortensen et al., 2018). Then, to get an IPA transcriptions of the English entries, the *English-to-IPA* library in Python (version 0.21) was used (Phillips, 2019).[79] This library uses the *Carnegie Mellon University (CMU) Pronouncing Dictionary* in order to convert English text into IPA. This tool was chosen for English because, as noted in the Epitran documentation, sound-symbol correspondence is so low in English that effective G2P systems in the language currently rely on pronouncing dictionaries.[80]

29 English entries (2.06% of 1,411) were removed because there was no IPA transcription for one or more of the words in the entry, since they could not be found in the CMU pronouncing dictionary. This included primarily phrases that are sometimes written using two separate words which in this case were conjoined (e.g., *shoulderblade*, *lowtide*, and *doorpost*), though some of the excluded words are more frequently written in a conjoined manner (e.g., *beeswax*, *stingray*, and *earlobe*),[81] and there were also some words that did not fit this pattern (e.g., *adze*, *defecate* and *nape*).

In addition, there was a problem with some of the characters used in the English the pronunciation dictionary. Specifically, these characters were inconsistent with the character set used by Epitran, which caused inconsistencies with the IPA transcriptions between the L1s and English. This can lead to issues in various ways, such as when it comes to calculations of lexical

[79] The relevant function was run so that only a single transcription was provided per word, in keeping with the other languages, though some words may have more than a single transcription available in the dictionary.

[80] Epitran itself currently relies on a similar dependency for its English G2P system (namely, the CMU Flite speech synthesis system) but the implementation of the dedicated English-to-IPA library was found to be more convenient from a practical perspective.

[81] Because the pronunciation of these phrases as a single word could be different from their pronunciation as separate units, the decision was made to remove them together with the other entry that had no IPA pronunciation listed, rather than derive their transcription based on the IPA transcription of their constituents.

distance and which was potentially problematic for calculations of lexical distance, since such characters are not recognized by the software used to calculate feature distance (PanPhon, which is explained in more detail later, in the section explaining how lexical distance was calculated).[82] This included the following:

- /g/ (Latin Small Letter G, U+0067, from the ‘Basic Latin’ block), which was replaced by /ɡ/ (Latin Small Letter Script G, U+0261, from the ‘IPA extensions block’).[83]
- The affricates /ʧ/ and /ʤ/, which were replaced by /t͡ʃ/ and /d͡ʒ/ respectively.

Finally, an important caveat is that Epitran warns that the transcriptions for several languages, including Portuguese and French in the present sample, should be approached with caution, since “It is not possible to provide highly accurate support for these language-script pairs due to the high degree of ambiguity inherent in the orthographies” (Mortensen, 2019).[84] This is important to take into account in any analyses involving these languages. In the present study, the lexical distance data for these languages was validated by comparing the correlations between the lexical distances in the parallel dictionaries to those in the Swadesh lists, for words that appear in both datasets, separately for each language. These analyses are shown in §7.3.3, which compares the two datasets, and overall, the correlations were high for all L1s, and though the correlations for those two languages were slightly lower than for the other languages, this difference was not statistically significant.

7.3.3 Comparison of lexical-distance datasets

We compared the lexical distances calculated based on the Swadesh lists with those calculated based on the parallel dictionaries, to see to what degree they correspond. A primary goal of this was to see if the data for French and Portuguese corresponds between the two datasets as much as the data for the other shared languages (German, Italian, and Spanish), since Epitran, which is the software used to derive phonological transcriptions from orthographical ones in the parallel dictionariones in the present dataset, warns that those languages are more likely to

[82] The problem characters were identified by collecting the set of all characters used in the English IPA transcriptions in the present sample, and calculating the feature edit distance between them and an empty string (such distance is explained later); cases where there was a distance of zero indicated that the character was not recognized by PanPhon.

[83] Character codes were checked on https://unicode-table.com/.

[84] An attempt was made to use pronouncing dictionaries for these languages too. However, the dictionaries that were found all had lower coverage, and while they may have been more accurate in some cases, there were also cases where they had identical transcriptions as the Epitran versions (or transcriptions that were highly similar), as well as cases where they introduced new issues. Because of this, and because of the added complexity that they lead to, it was decided to use only the Epitran versions for the languages.

involve errors, due to ambiguity inherent in their orthogrphy. In addition to issues that arise due to the phonological transcription proess, differences between the datasets could be attributed to amany other causes, such as dialectical variation in the chosen transcription.

We calculated the correlation between the phonological LDN in the two datasets, for the 35 English meanings per L1 that appeared in both datasets initially, including multi-word entries (175 entries total). The overall correlation between the two datasets was $r = .738$ (*95% CI* = .662–.799, $p < .001$), and correlations for each L1 are listed in Table 20.[85]

Table 20. Correlation between phonological LDN in the Swadesh lists and parallel dictionaries, with $n = 35$ for each language.

Language	*Pearson's r*	*95% CI*	*p*
French	.637	.385-.800	< .001
German	.731	.526-.856	< .001
Italian	.761	.573-.873	< .001
Portuguese	.615	.355-.787	< .001
Spanish	.828	.683-.910	< .001

Portuguese and French had the lowest correlations, but these correlations were still high, and the differences in correlation between these languages and the others was not statistically significant, as shown in Table 21.

[85] The correlation in distances between single-word entries that appear in both datasets (22 entries per L1, 110 total), was similar: $r(108) = .681$ ($p < .001$, *95% CI* = [.566, .770]).

Table 21. Difference in correlation between the languages. Specifically, Portuguese and French are compared to Spanish (the language with the highest correlation) and German (the language with the lowest correlation aside from the first two).

	German	**Spanish**
Portuguese	*Fisher's z* = -0.8558, *p* = .3921 *Zou's 95% CI* = -0.4047, 0.1516]	*Fisher's z* = -1.8516, *p* = .0641 *Zou's 95% CI* = [-0.4854, 0.0126]
French	*Fisher's z* = -0.7164, *p* = .4737 *Zou's 95% CI* = [-0.3755, 0.1675]	*Fisher's z* = -1.7123, *p* = .0868 *Zou's 95% CI* = [-0.4558, 0.0275]

Note. Differences were calculated using the *cocor* package in R (Diedenhofen & Musch, 2015), with the correlation for each language being treated as an independent group.

Although the differences are close to significance in the case of the comparison with Spanish, they are far from it in the comparison with German. As such, and given that the correlations across datasets for Portuguese and French are high overall, it was decided to keep these languages in the present sample, with the caveat that their phonological transcriptions likely contain slightly more "noise" than those of the other languages.

7.3.4 Word frequency information

Frequency was calculated using the *wordfreq* library in Python (Speer et al., 2018). The *wordfreq* library was chosen for several reasons, including the large dataset that it uses, the diverse sources that its dataset is based on, and its robust process for calculating frequency. Specifically, when it comes to the data that this library is based on:

> This data comes from a Luminoso project called Exquisite Corpus, whose goal is to download good, varied, multilingual corpus data, process it appropriately, and combine it into unified resources such as wordfreq.
>
> Exquisite Corpus compiles 8 different domains of text, some of which themselves come from multiple sources:
>
> - **Wikipedia**, representing encyclopedic text
> - **Subtitles**, from OPUS OpenSubtitles 2018 and SUBTLEX
> - **News**, from NewsCrawl 2014 and GlobalVoices
> - **Books**, from Google Books Ngrams 2012

- **Web text**, from ParaCrawl, the Leeds Internet Corpus, and the MOKK Hungarian Webcorpus
- **Twitter**, representing short-form social media
- **Reddit**, representing potentially longer Internet comments
- **Miscellaneous** word frequencies: in Chinese, we import a free wordlist that comes with the Jieba word segmenter, whose provenance we don't really know

(Speer, 2020)

In particular, the frequency data for the languages that were included in the present study comes from the sources outlined in Table 22.

Table 22. Sources for the data on the word frequencies in the languages that were examined in the present study, based on the information in Speer (2020).

Language	Number of Sources	Wiki	Subs	News	Books	Web	Twitter	Reddit	Misc.
English	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
French	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
German	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
Italian	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
Portuguese	5	Yes	Yes	Yes	-	Yes	Yes	-	-
Spanish	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-

As such, the languages had similar data sources, and all of them had sufficient data to be designated as having a ‘large’ wordlist, meaning that they cover words that “appear at least once per 100 million words”, in contrast with ‘small’ wordlists, which “cover words that appear at least once per million words” (Speer, 2020). However, note that in the present study we only used the frequency data for English, since the goal of using it was to simply control for the rate of usage of the target words in English

The *wordfreq* library combines the frequencies from different sources in a way that is meant to minimize the impact of potential outliers.[86,87] This is referred to as the *figure-skating metric*, after the scoring system of Olympic figure skating, and involves the following:

- Find the frequency of each word according to each data source.
- For each word, drop the sources that give it the highest and lowest frequency.
- Average the remaining frequencies.
- Rescale the resulting frequency list to add up to 1.

(Speer, 2020)

In the present study, we focus on the log-frequency of words, rather than on frequency, given the large body of research suggesting that log-frequency is often better associated with linguistic outcomes (Baayen, 2001; Baayen et al., 2006; Balota et al., 2001; J. B. Carroll, 1967; Davis, 2005; de Groot & Keijzer, 2000; Gordon & Caramazza, 1982; Jurafsky, 2003; Kuperman et al., 2012; Rubenstein & Pollack, 1963; Scarborough et al., 1977; Segui et al., 1982; Tanaka-Ishii & Terada, 2011; Winter, 2019), including in the context of similar L2 outcomes as we examine here (Bosma et al., 2019; Carrasco-Ortiz et al., 2021; De Wilde et al., 2020, 2021; Otwinowska et al., 2020; Otwinowska & Szewczyk, 2019; Poort & Rodd, 2017; Sadat et al., 2016; van de Ven et al., 2019).

Specifically, the *wordfreq* library offers an improved log-frequency measure, called *Zipf frequency*, which is also used in several other studies on similar L2 outcomes (Bosma et

[86] Though not crucial for the present analysis, it is worthwhile to point out that wordfreq relies on frequency bins for performance reasons. As stated on the project page: “wordfreq's wordlists are designed to load quickly and take up little space in the repository. We accomplish this by avoiding meaningless precision and packing the words into frequency bins. In wordfreq, all words that have the same Zipf frequency rounded to the nearest hundredth have the same frequency. We don't store any more precision than that. So instead of having to store that the frequency of a word is .000011748975549395302, where most of those digits are meaningless, we just store the frequency bins and the words they contain. Because the Zipf scale is a logarithmic scale, this preserves the same relative precision no matter how far down you are in the word list. The frequency of any word is precise to within 1%. (This is not a claim about accuracy, but about precision. We believe that the way we use multiple data sources and discard outliers makes wordfreq a more accurate measurement of the way these words are really used in written language, but it's unclear how one would measure this accuracy.)” (Speer, 2020).

[87] Though this is not relevant for the present study, which focused on single-word entries, the *wordfreq* library is capable of estimating the frequency of multi-word combinations, based on the frequency of their constituents, where “The word frequencies are combined with the half-harmonic-mean function in order to provide an estimate of what their combined frequency would be” (Speer, 2020). However, an important caveat about this approach is the following: “This method of combining word frequencies implicitly assumes that you're asking about words that frequently appear together. It's not multiplying the frequencies, because that would assume they are statistically unrelated. So if you give it an uncommon combination of tokens, it will hugely over-estimate their frequency.” (Speer, 2020).

al., 2019; Carrasco-Ortiz et al., 2021; De Wilde et al., 2020, 2021). This measure, which was developed by van Heuven et al. (2014), is described as follows in the wordfreq library:

> The Zipf frequency of a word is the base-10 logarithm of the number of times it appears per billion words. A word with Zipf value 6 appears once per thousand words, for example, and a word with Zipf value 3 appears once per million words.
>
> Reasonable Zipf values are between 0 and 8, but because of the cutoffs described above, the minimum Zipf value appearing in these lists is 1.0 for the 'large' wordlists and 3.0 for 'small'. We use 0 as the default Zipf value for words that do not appear in the given wordlist, although it should mean one occurrence per billion words.
>
> (Speer, 2020)

This measure was developed with the goal of creating a standardized measure, that would be independent of corpus size, and that would fulfill the following conditions:

> (1) It should be a logarithmic scale (e.g., like the decibel scale of sound loudness).
>
> (2) It should have relatively few points, without negative values (e.g., like a typical Likert rating scale, from 1 to 7).
>
> (3) The middle of the scale should separate the low-frequency words from the high-frequency words.
>
> (4) The scale should have a straightforward unit.
>
> (van Heuven et al., 2014, p. 1179)

To understand this scale, consider the examples in Table 23.

Table 23. Examples of words with different Zipf values, taken from van Heuven et al. (2014, p. 1180).

Zipf value	Frequency per million words	Examples of words
1	0.01	bioengineering, farsighted, harelip
2	0.1	airstream, doorkeeper, neckwear
3	1	cornerstone, dumpling, perpetrator
4	10	dirt, muffin, widespread
5	100	basically, drive, spot
6	1,000	great, other, years
7	10,000	and, have, I

Note. In the original study, words with a Zipf value of 3 and lower were considered low-frequency words, while words with a Zipf value of 4 and higher were considered high-frequency words.

7.4 Appendix D: Supplementary information for study 1 (on lexical diversity)

7.4.1 Lexical distance

[This first three sub-sections in this section largely overlap with the corresponding material of the supplementary information for study 2, though the latter has some more information in the part on the validation of lexical distance. As such, and due to space constraints, I removed the overlapping content from this section; for the relevant material, see the first section of Appendix E.

Included below is the one sub-section that appears in this study but not the other, since it focuses on using the Swadesh lists for calculating distance at the language—rather than word—level. Nevertheless, some of this material is also mentioned in parts pertaining to the second study, and primarily Appendix C, which contains information on the calculation of lexical distance.]

7.4.1.1 Validation of ASJP-based Swadesh lists

The previous sub-section outlines the support for Levenshtein distance as a measure of language distance, both in general and in the context of SLA. In addition, below we outline the support for the use of the ASJP Swadesh lists as a source for calculating lexical distance.

First, the key evidence in support of these lists comes from studies that cross-validated Levenshtein distances that were calculated from these lists, by comparing them with other measures of language distance, and showing that they strongly correlate. This includes correlations between the ASJP-based distances and:

- Expert-based cognancy judgments, on historical-comparative grounds (Schepens, van der Slik, et al., 2013b).
- Psychoacoustic, psycholinguistic, phonetic, and phonological measures of distance (Brown et al., 2013).
- Taxonomic distances in the *World Atlas of Language Structures* (WALS) and the *Ethnologue* (Holman et al., 2008b).
- Distances based on morphological features in the WALS (Schepens, van der Slik, et al., 2013a) and typological features (Bakker et al., 2009).

In addition, further support for these lists comes from studies that used it for language classification and linguistic phylogenetics, and involved various forms of validity and reliability checks (Brown et al., 2008; Holman et al., 2008b, 2008a; Pompei et al., 2011; Wichmann, 2019; Wichmann et al., 2010).[88]

Finally, distances from these lists were also used successfully in the context of SLA, to predict L2 Dutch proficiency, in a sample of 35 Indo-European L1s (Schepens, van der Slik, et al., 2013b), and in a sample of 39 Indo-European and 34 non-Indo-European L1s (Schepens, van der Slik, et al., 2013a). This aligns with similar findings on L2 Dutch proficiency, which are based on two different types of cognancy judgments (Schepens et al., 2020; van der Slik, 2010).

However, despite this support for the use of the ASJP lists, it is important to also note their limitations. First, the ASJP uses a specialized set of characters in its transcription (outlined in Brown et al., 2008), which collapses certain phonological distinctions to facilitate transcription, by transcribing different segments using the same character (e.g., the IPA: /b/ and /β/ are both transcribed using the ASJP /b/). Second, there may be various issues with the ASJP transcriptions, meaning that some words may be transcribed incorrectly, and that the wrong words be may selected for transcription in some cases. As such, distances that are based on these lists should preferably be used for large-scale analyses that can accommodate this (as in the present study). Furthermore, distances based on these lists should be interpreted with caution, and compared with distances from other sources where possible (again, as in the present study).

In summary, it is important to be cautious when relying on distances calculated based the ASJP lists, just as it is important to be cautious about distances calculated using other sources. This also means that it is beneficial to replicate analyses using distances from different sources, as we do in the present study using a binary measure of language distance. Nevertheless, these is extensive support for the use of these lists, as outlined above, so we believe that the use of the list here is reasonable for our purposes.

[88] The ASJP lists have also been used for this purpose with language measures other than Levenshtein distance. For example, this includes studies that used *point-wise mutual information* (PMI) as the distance measure, although they also used Levenshtein distance as a first step to identify potential cognancy (Jäger, 2015, 2018).

7.4.2 Model diagnostics (assumption checks)

Figures 16 and 17 contain the assumption checks for the mixed-effects models that appeared in the study, and namely for linearity, homoscedasticity, normality of residuals, normality of random effects, lack of collinearity, and lack of influential observations (Hox et al., 2018; Winter, 2019). The plots were generated using the dedicated *performance* package in the *easystats* ecosystem in R (Lüdecke et al., 2021).

When interpreting these diagnostic plots, we follow two notable recommendations from Winter's (2019) relevant work, and namely the focus on primarily visual techniques for diagnostic purposes, and the use of assumption checking on as a way to determine whether there are any major issues with the model. As Winter notes in this regard:

> Newcomers to regression modeling often find it discomforting that the assumptions are assessed visually. In fact, formal tests for checking assumptions do exist, such as the Shapiro-Wilk test of normality. However, applied statisticians generally prefer visual diagnostics (Quinn & Keough, 2002; Faraway, 2005, 2006: 14; Zuur et al., 2009, Zuur, Ieno, & Elphick, 2010). The most important reason for using graphical validation of assumptions is that it tells you more about your model and the data. [Footnote 7: Here are some other reasons: each of these tests also has assumptions (which may or may not be violated), the tests rely on hard cut-offs such as significance tests (even though adherence to assumptions is a graded notion), and the tests may commit Type I errors (false positives) or Type II errors (false negatives) (see Chapter 10 for an explanation of these concepts).] For example, the residuals may reveal a hidden nonlinearity, which would suggest adding a nonlinear term to your model (see Chapter 8). Or the residuals may reveal extreme values that are worth inspecting in more detail. One should also remember that a model's adherence to the normality and constant variance assumptions is not a strict either/or. Faraway (2006: 14) says that 'It is virtually impossible to verify that a given model is exactly correct. The purpose of the diagnostics is more to check whether the model is not grossly wrong.'"
>
> (Winter, 2019, pp. 109-110)

Based on this approach, and based on the figures below, there do not appear to be issues with the model that are substantial enough to invalidate its use. Namely, the mild heteroscedasticity is unlikely to be an issue, especially given that the potential consequence of having standard errors that are too small (Hox et al., 2018, p. 248) would not substantially change the findings of the study, due to the null results for lexical distance and its interaction with proficiency, and also the robustness of the effects of L2 proficiency. In addition, note that the findings in the mixed-models are also supported by the other analyses in the main paper, and specifically the estimated marginal means and the linear models per CEFR level.

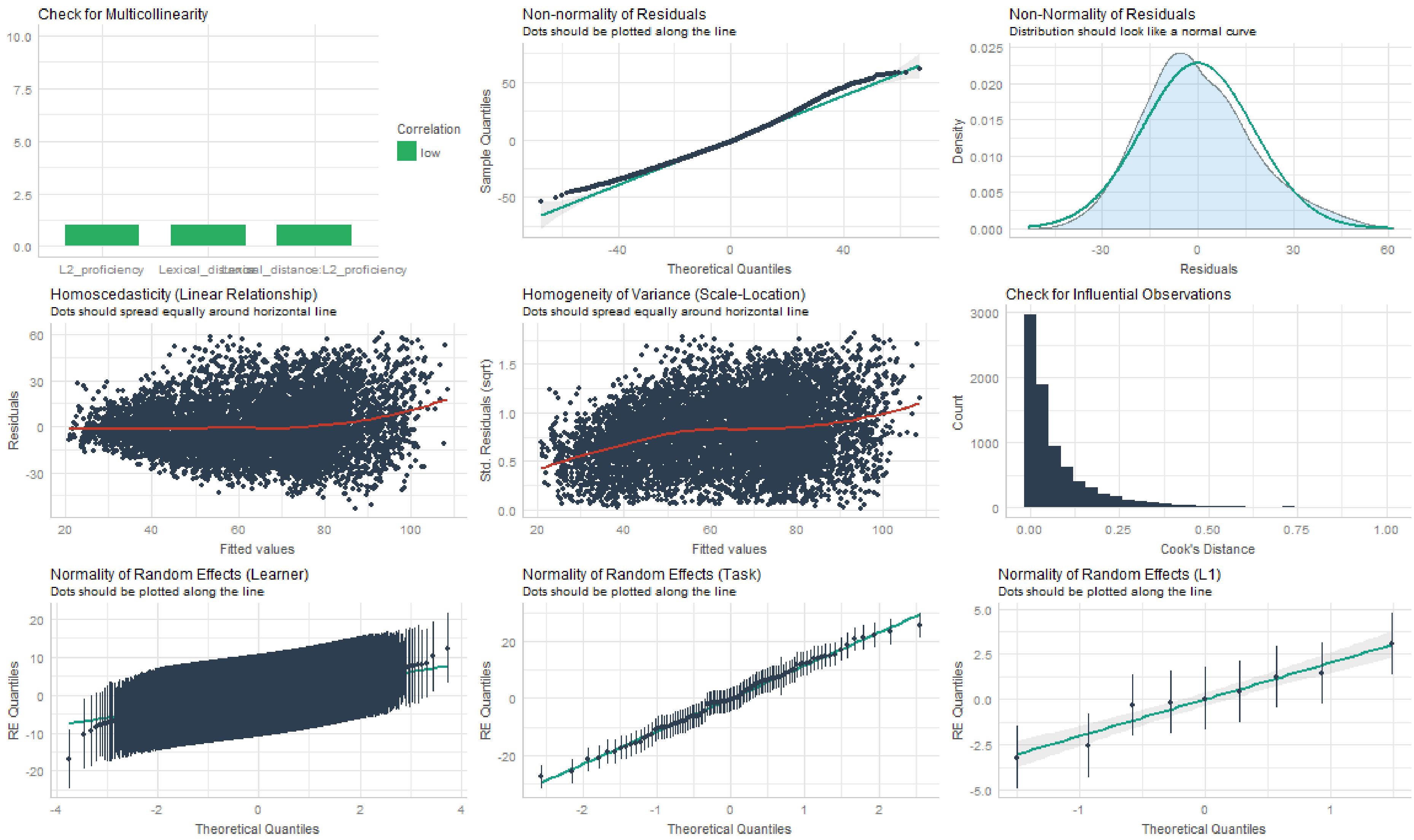


Figure 16. Assumption checks for the main models in the first corpus.

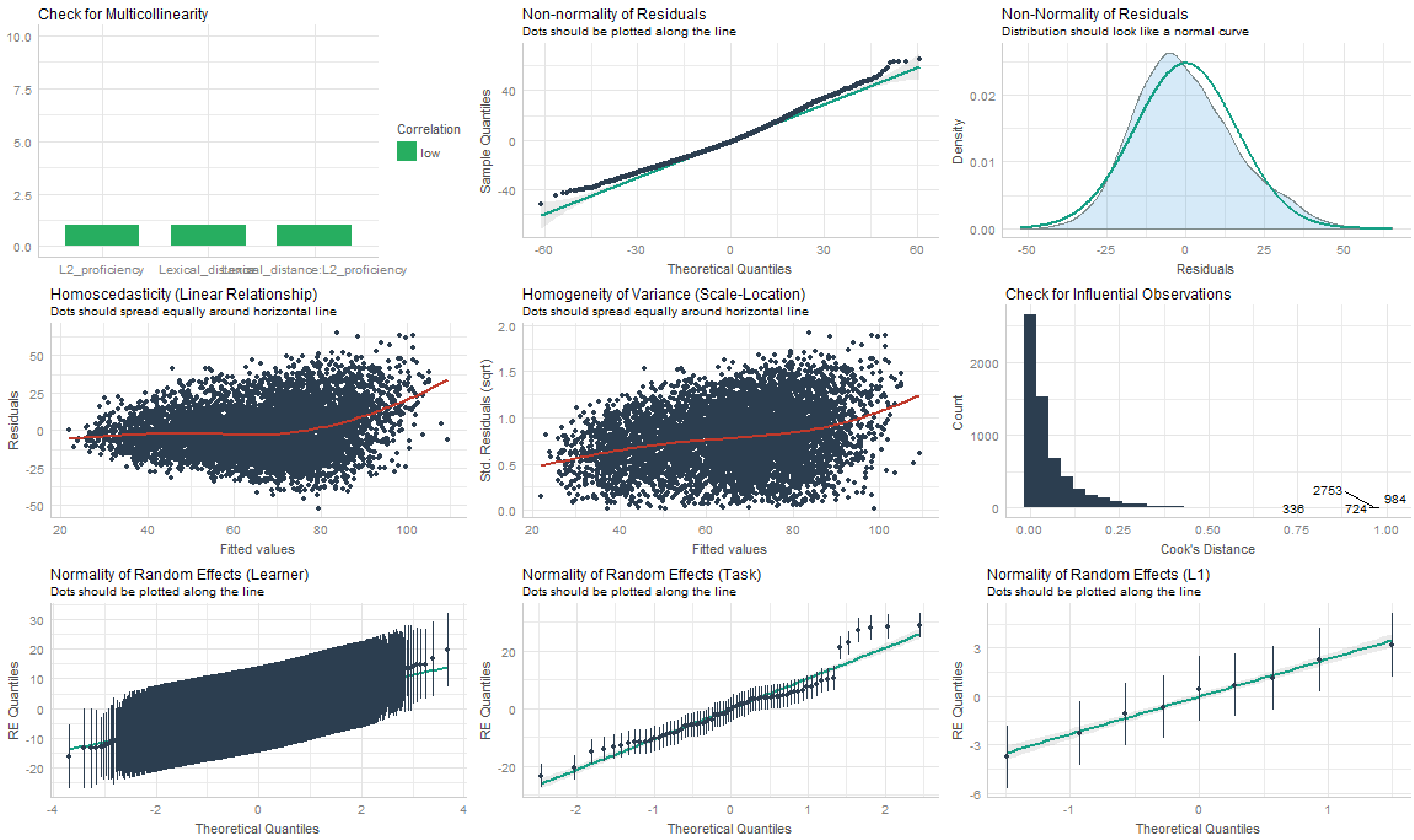


Figure 17. Assumption checks for the main models in the second corpus.

7.4.3 Information on MTLD

7.4.3.1 Technical approach for calculating MTLD (including spelling correction)

In our analyses, we used the MTLD scores that come pre-calculated with the EFCAMDAT Cleaned Subcorpus (Shatz, 2020). These scores were calculated based on the spelling-corrected version of each text, generated using the *autocorrect* library in Python (McCallum, 2019), since spelling errors can artificially inflate the number of types in the text (Granger & Wynne, 1999). They were calculated using the *lexical-diversity* library in Python (Kyle, 2018), first using the library's *tokenize* function to tokenize each text, and then using the *mtld* function to calculate the MTLD.

There are two main caveats about this approach. First, if the autocorrect function—for which there is no validation information available—corrects misspelled words by transforming them into a different word than learners intended, this could still result in some inflation of lexical diversity scores. Second, it is possible to add further steps to the pre-processing of the texts—especially lemmatization—to potentially increase the validity of the resulting lexical-diversity calculations.

As such, we also calculated MTLD scores using the recently released *TAALED* Python library (Kyle et al., 2021),[89] which utilizes more sophisticated pre-processing for texts.[90] Specifically, TAALED involves multiple steps of preprocessing: tokenizing the text, removing most punctuation, adding part-of-speech tags (to disambiguate homographs), lemmatizing each word, checking for misspelled words (which are then ignored, rather than corrected), and converting all words to lower-case.

In practice, though this may be beneficial, it does not substantially alter our conclusions. Specifically, the correlation between the original MTLD scores and the new (TAALED-based ones) was very strong: r = .77, p < .001, *95%CI* = [.76, .78] in the first corpus and r = .74, p < .001, *95%CI* = [.73, .75] in the second (in the samples used in the main models, after the

[89] We did not use this library in the first place because it was released very recently, after we concluded the main analyses for this study. In addition, we did not switch to using it from the main analyses, since it is still in beta, so it has not yet been as extensively tested and used by others as the simpler approach, which also means that the MTLD values calculated using its approach are likely to be less comparable with those in other studies. Nevertheless, as shown next, the results based on these scores closely mirror those from the main models, so this is not crucial for our study. Also, note that all the MTLD values that we used available for further analysis, together with the rest of our data, in the study's OSF repository.

[90] We used version 0.22 of TAALED. Note that this library is developed by the same person as the *lexical-diversity* library which we used to calculate the main MTLD scores, and utilizes the same underlying functions for calculating lexical diversity. In addition, the pre-processing involved with this library relies on the *pylats* (Kyle, 2022) and *spaCy* (Honnibal et al., 2020) Python libraries; versions 0.24 and 3.2.4 were used respectively.

removal of the outliers). In addition, as shown in Table 24 (Figures 18 and 19 contain the associated assumption checks), our key findings replicate when using the new MTLD scores; this includes the null effect of lexical distance and of its interaction with L2 proficiency, and the strong effects of *L2 proficiency* and *task*. This means that our conclusions remain the same regardless of which of these two approaches we use in our analyses.

Table 24. Results of the mixed-effects linear regression models, with lexical diversity (MTLD) as the response variable; here, the MTLD are those calculated using the *TAALED* library in Python. Under fixed effects, *lexical_distance* is the mean lexical distance between the L1 and English (0–1), and *L2_proficiency* is the EFCAMDAT level associated with each text (1–12). In the statistics, *std. B* and *std. 95% CI* provide information on the standardized coefficients, which were calculated by refitting the model on standardized data. Under random effects, σ^2 denotes the residual variance, τ_{00} denotes between-subjects (or groups) variance, *ICC* denotes the intraclass correlation coefficient, and *N* denotes the number of data points within each sampling unit. Finally, *observations* denotes the total number of texts in each sample, *Mar. [Marginal] R^2* denotes the proportion of the variance described by the fixed effects, and *Cond. [Conditional] R^2* denotes the proportion of the variance described by both the fixed and random effects.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	50.98	1.00	49.02 – 52.94	<0.001	0.01	-0.07 – 0.09	52.00	1.03	49.99 – 54.01	<0.001	0.00	-0.09 – 0.09
Lexical_distance	-5.30	5.88	-16.83 – 6.23	0.368	-0.02	-0.05 – 0.02	0.89	5.62	-10.13 – 11.91	0.874	0.00	-0.03 – 0.04
L2_proficiency	3.26	0.27	2.73 – 3.78	<0.001	0.46	0.38 – 0.53	2.69	0.28	2.15 – 3.24	<0.001	0.42	0.33 – 0.50
Lexical_distance * L2_proficiency	0.79	0.89	-0.96 – 2.54	0.378	0.01	-0.01 – 0.03	0.57	0.96	-1.31 – 2.46	0.550	0.01	-0.02 – 0.03
Random Effects												
σ^2	366.84						302.24					
τ_{00}	26.10 Learner						44.73 Learner					
	76.72 Task						62.17 Task					
	1.25 L1						1.01 L1					
ICC	0.22						0.26					
N	9 L1						9 L1					
	5385 Learner						4357 Learner					
	95 Task						71 Task					
Observations	8081						6129					
Mar. R^2 / Con. R^2	0.209 / 0.384						0.175 / 0.392					

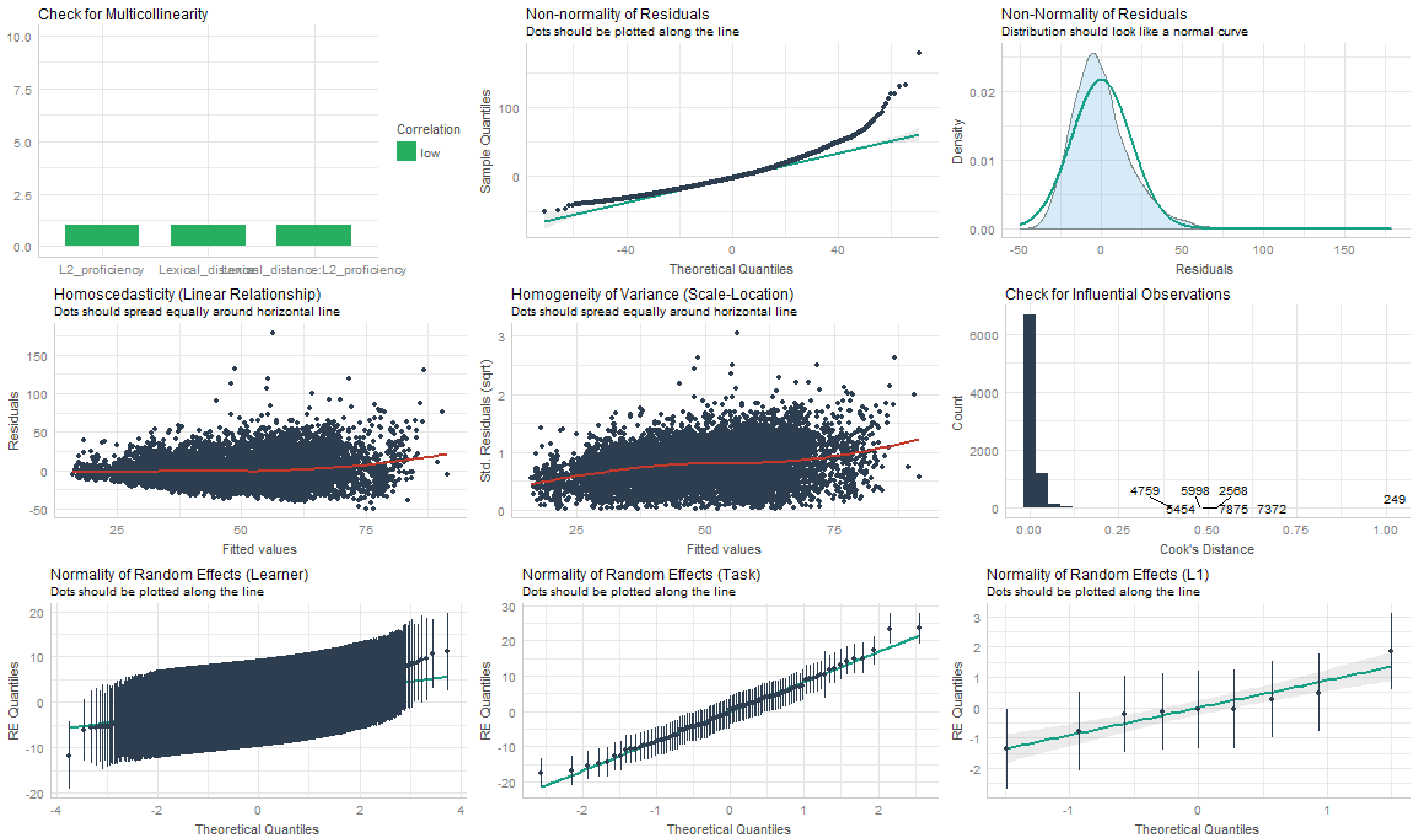


Figure 18. Assumption checks for the TAALED-based models in the first corpus.

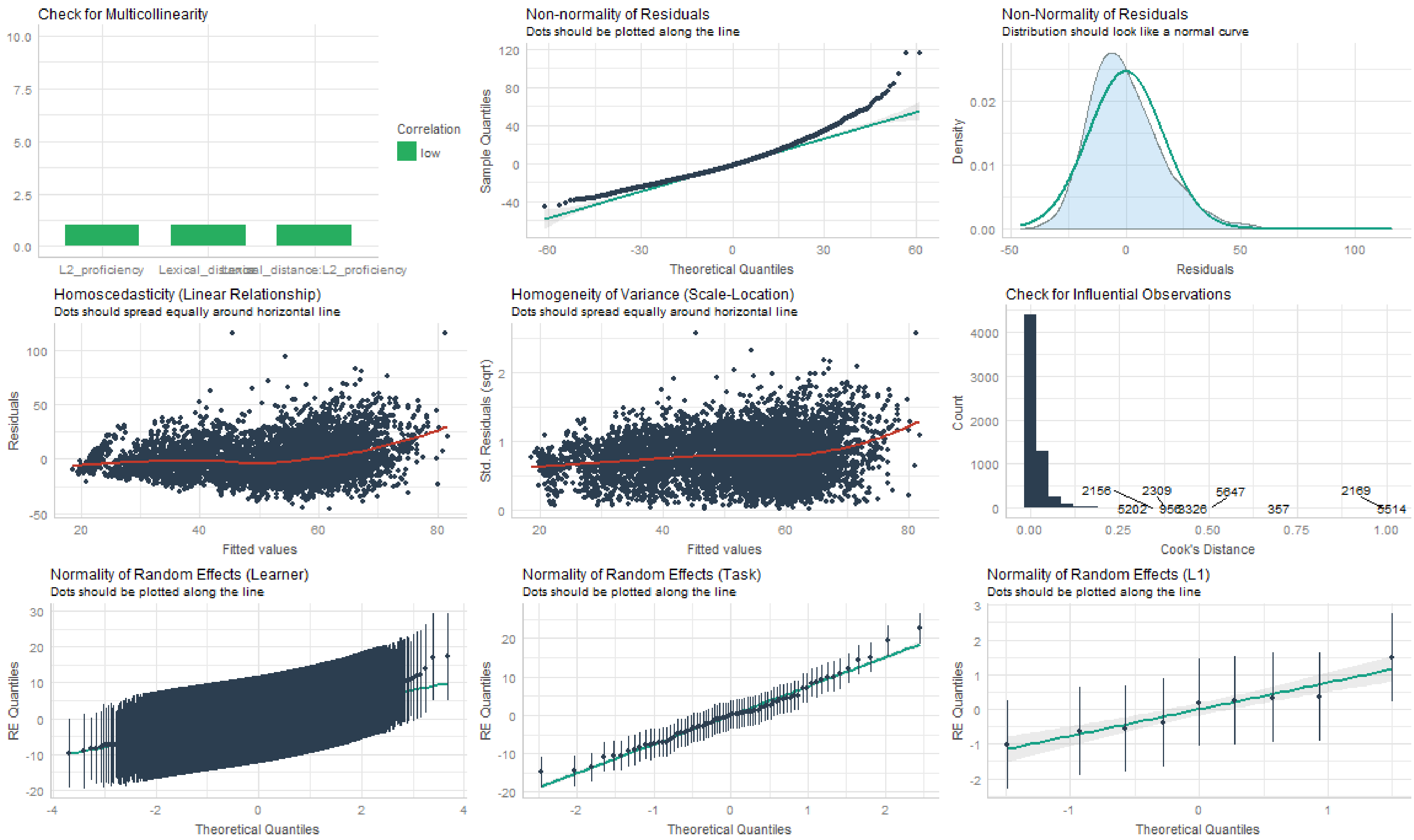


Figure 19. Assumption checks for the TAALED-based models in the second corpus.

7.4.3.2 MTLD outliers

The study used MTLD to assess learners' lexical diversity, where higher MTLD indicates greater lexical diversity. Because MTLD varies substantially based on various factors, such as *L2 proficiency* and *task*, there is no simple way to characterize what general mean values of MTLD are expected to be (Alexopoulou et al., 2017; Hout & Vermeer, 2010; Jarvis, 2013; Kojima & Yamashita, 2014; Mazgutova & Kormos, 2015; Murakami & Alexopoulou, 2016; Torruella & Capsada, 2013; Treffers-Daller, 2013; Treffers-Daller et al., 2018; Yu, 2010; Zenker & Kyle, 2021). Nevertheless, a very rough characterization, based on the aforementioned studies, is that mean MTLD values tend to fall between 50–100, although both mean MTLD and the MTLD of individual texts can be lower or higher than that, in approximately the 10–150 range.

In the present study, before running the main analyses, we identified and removed texts that were classified as outliers in terms of their lexical diversity. The goal of this was to remove extreme values that were likely to be caused by technical issues with the assessment of MTLD, particularly when it comes to short texts (see the section on text length and MTLD for more information on this).

The first step to this was to remove texts with an anomalous *absolute* MTLD, and specifically an MTLD of 0. Such MTLD represents a unique case where there is absolutely no lexical repetition in the text, so the TTR is 1 throughout, leading to an MTLD of 0. These cases primarily reflect the technical limitations of lexical diversity measures when it comes to short texts with a list-like structure.[91] As such, this MTLD value was only found in short texts, with wordcounts of 20–39 in the first corpus (mean = 27.78, SD = 5.21) and 20–34 in the second corpus (mean = 22.85, SD = 3.83) (see the section on text length and MTLD for more information on this). Furthermore, there was a substantial gap in MTLD between these texts and texts with a non-zero MTLD, as the minimal non-zero MTLD was 8.01 in the first corpus and 11.12 in the second.

Then, additional outliers were identified and removed based on anomalous *relative* MTLD, using *Tukey's method* (Hoaglin et al., 1986; Kannan et al., 2015; Rousseeuw & Croux, 1993; Tukey, 1977). Specifically, outlier texts were defined as texts that had an MTLD at least 1.5 *interquartile ranges* (IQR) below the 1st quartile of MTLD or above the 3rd quartile, either

[91] An example of such an outlier text is the following: "Dear Mr hang, there are 33 computers,5 pens,9 keyboards,3 headphones,6 boxes, 46 desks and chairs in the office, a desk of them have broken, need to repair thank you Dora" (text #614876 in the first corpus, lesson #2).

within their task or within their EFCAMDAT proficiency level. It was necessary to remove these texts because their extreme MTLD values appeared to be primarily reflective of technical issues with the assessment of lexical diversity, as many of them had MTLD values that are far above what could be expected at that proficiency level, or far above what would be expected in general, on even complex native-level texts. For example, one text in task #22 in the first corpus had an MTLD of 493.92, while the median MTLD at the associated EFCAMDAT proficiency level was 43.56, and the median MTLD at the highest proficiency level in the corpus (corresponding to the top of the CEFR B2 range) was 81.78.[92] It was necessary to account for both *task* and *level* median MTLD, since MTLD is often more consistent within tasks due to task effects, but individual tasks are also more likely to be influenced by extreme outliers than whole levels, particularly when the task prompt elicited writing that was likely to lead to extreme MTLD due to technical issues. Specifically, this was primarily the case in tasks that elicited short, list-like texts, such as task #22 in the first corpus ("Writing your online profile"), which was responsible for 14.32% of outlier texts in the first corpus, but 0% of outliers in the second corpus. The task had many extreme outliers in terms of high MTLD, so the median MTLD for the task (after removing zero-MTLD outliers) was 109.76, while the median MTLD for the EFCAMDAT proficiency level in which this task appeared (level 3) was 43.56, and the median MTLD at the highest proficiency level in the corpus (level 12) was 81.78.

Overall, the proportion of texts that were designated as outliers was small and similar in the two samples (4%–5%). All the texts that were removed as outliers appear in the study's OSF repository, and relevant statistics regarding outlier texts are shown in Table 25. After the removal of outliers, there were 8,081 texts in the first corpus and 6,129 in the second.

In addition, Table 26 shows the results of supporting models that were calculated on the full samples, before the removal of outliers (Figures 20 and 21 show the associated assumption checks). As expected, there were more issues with these models' assumptions compared to the main models (e.g., with the non-normality of the residuals and with heteroscedasticity), which provides support for the removal of the outliers. Nevertheless, the findings in these models largely mirror those of the main models, in the sense that the lexical-

[92] The text was "Name: Julio Age: 32 birthday: 06 July Like doing: cooking, listening to music, chatting on line, surfing the internet. Lives in: Sao Paulo, Brazil. favorite Season: summer animal: dog Time: evening Day: Sunday Number: 6 I can cook, but I cant draw." (text #66095 in the first corpus).

distance predictor and the interaction term were non-significant and negligible, whereas the L2 proficiency predictor and the task effects were significant and substantial.

Table 25. Descriptive statistics of the outlier texts.

	First Corpus	Second corpus
Number of L1s in sample	9	9
Number of texts (total in sample)	8500	6390
Number of outliers	419	261
Proportion of outliers (out of total texts)	.05	.04
Proportion of texts per L1 (in full samples) [a]	.11	.11
Min proportion of texts per L1 (in outliers)	.09	.07
Max proportion of texts per L1 (in outliers)	.15	.19
SD of proportion of texts per L1 (in outliers) [b]	.02	.04
Mean EFCAMDAT proficiency level (in full samples) [c]	6.47	6.49
Mean EFCAMDAT proficiency level (in outliers) [c]	4.67	5.21
Mean wordcount (in full samples)	82.59	90.21
Mean wordcount (in outliers)	63.70	75.00
Proportion of outliers with MTLD of '0'	.12	.05
Proportion of outliers with MTLD above the cutoff [d]	0.998	1.00
Proportion of outliers in only task MTLD [e]	.29	.25
Proportion of outliers in only level MTLD [e]	.35	.24
Proportion of outliers in both task and level MTLD [e]	.24	.47

[a] The *proportion of texts per L1* is equal to 1 divided by the number of L1s in the sample (9). The mean proportion of texts per L1 in the outliers is equal to the proportion of texts per L1 in the corresponding sample, as all L1s that appear in the original samples contained outliers.

[b] Since the full samples had an equal number of texts per L1 in the second corpus and an almost equal number of texts in the first corpus, the *SD* of the proportion of texts per L1 in the full samples was 0 in the second corpus and 0.001 in the first.

[c] This is on a scale of 1-12 in the present sample, where each 3 levels correspond to a single CEFR level (e.g., EFCAMDAT levels 1-3 correspond to CEFR A1).

[d] This refers to the proportion of outliers, out of all the outlier texts, that were removed because they had an MTLD greater than 1.5 IQRs above the 3rd quartile of MTLDs in their corresponding task or proficiency level, after the removal of texts with an MTLD equal to 0.

[e] Non-zero outliers were removed either because their MTLD exceeded the threshold for their task, for their proficiency level, or both. This removal took place after the removal of zero-MTLD outliers.

Table 26. Results of the models that were calculated on the full sample (before the removal of outliers).

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	67.99	1.62	64.82 – 71.16	<0.001	0.00	-0.09 – 0.09	69.50	1.89	65.79 – 73.20	<0.001	-0.00	-0.12 – 0.12
Lexical_distance	-5.91	8.93	-23.42 – 11.60	0.508	-0.01	-0.05 – 0.03	13.79	15.98	-17.52 – 45.11	0.388	0.03	-0.04 – 0.11
L2_proficiency	3.32	0.44	2.47 – 4.18	<0.001	0.34	0.25 – 0.42	2.81	0.44	1.96 – 3.67	<0.001	0.31	0.22 – 0.41
Lexical_distance * L2_proficiency	1.39	1.27	-1.10 – 3.89	0.274	0.01	-0.01 – 0.03	0.54	1.38	-2.16 – 3.24	0.694	0.00	-0.02 – 0.03
Random Effects												
σ^2	763.83						545.84					
τ_{00}	77.46 $_{\text{Learner}}$						172.63 $_{\text{Learner}}$					
	206.53 $_{\text{Task}}$						153.32 $_{\text{Task}}$					
	2.99 $_{\text{L1}}$						11.47 $_{\text{L1}}$					
ICC	0.27						0.38					
N	9 $_{\text{L1}}$						9 $_{\text{L1}}$					
	5615 $_{\text{Learner}}$						4512 $_{\text{Learner}}$					
	95 $_{\text{Task}}$						71 $_{\text{Task}}$					
Observations	8500						6390					
Mar. R^2 / Cond. R^2	0.112 / 0.354						0.099 / 0.443					

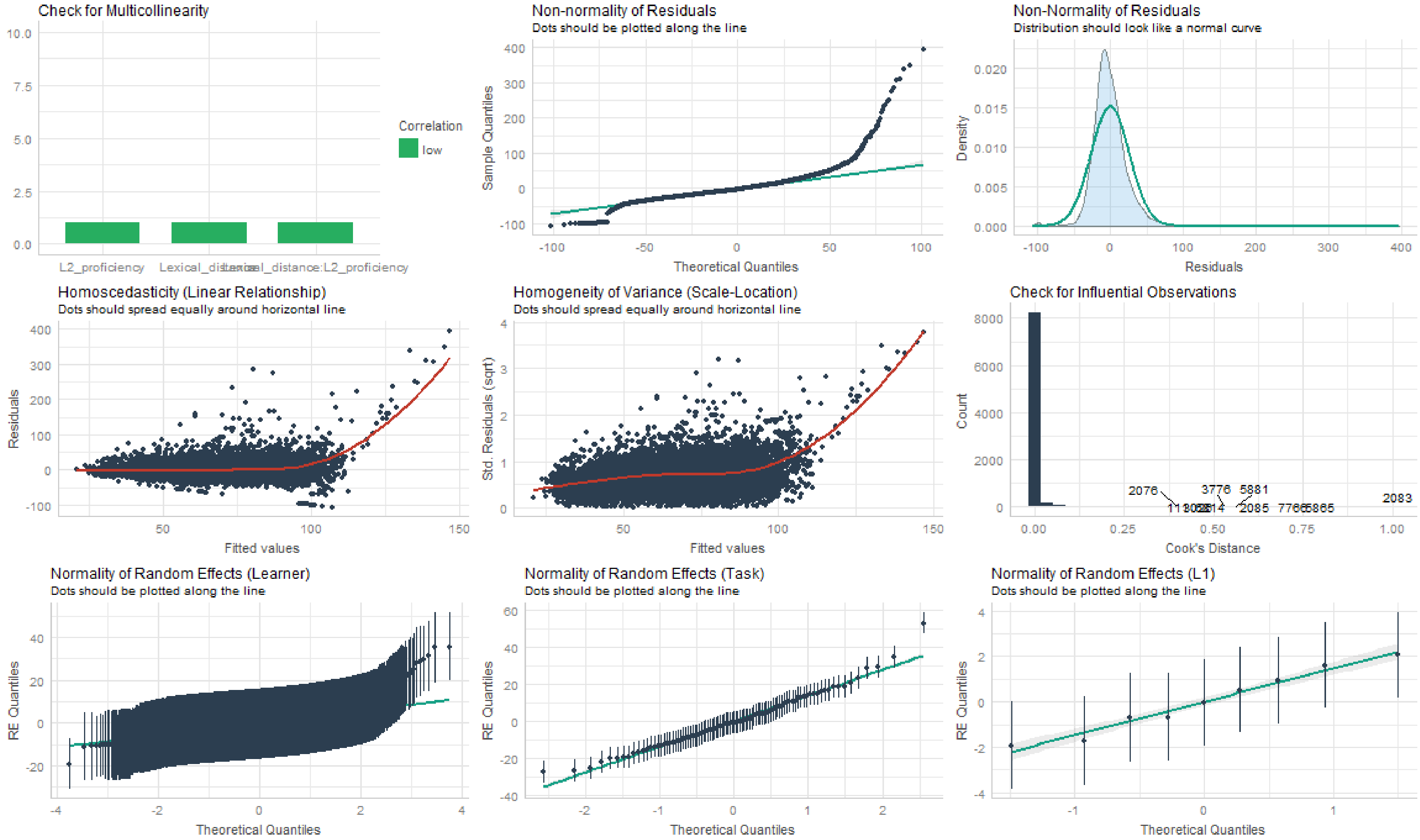


Figure 20. Assumption checks for the model before the removal of outliers, in the first corpus.

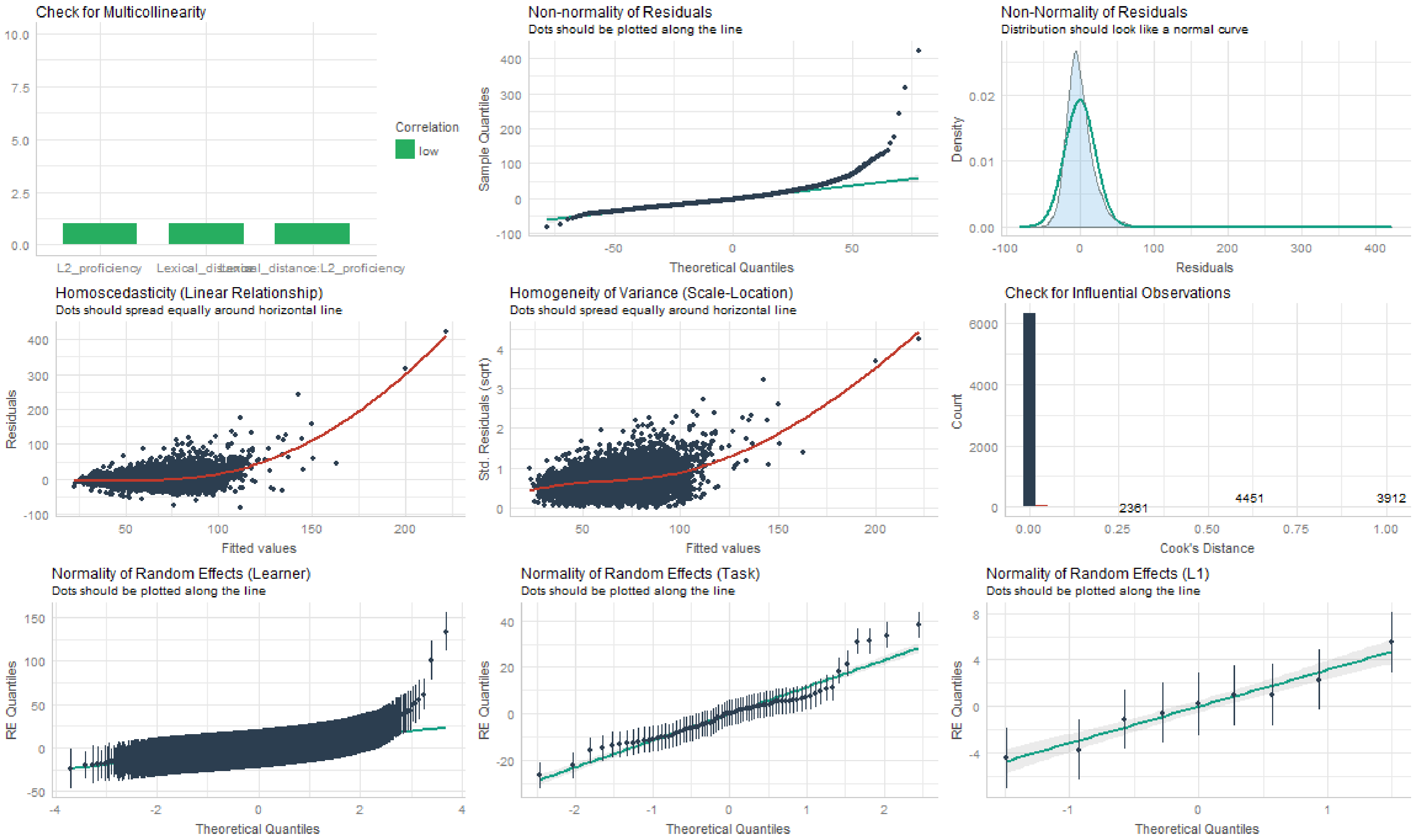


Figure 21. Assumption checks for the model before the removal of outliers, in the second corpus.

7.4.3.3 Text length (wordcount) and MTLD

Prior studies have shown that MTLD is relatively robust to short texts and to variations in text length, which means that it is less influenced by them than most other lexical diversity measures, with the possible exception of MATTR (Fergadiotis et al., 2015; Koizumi, 2012; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020; Zenker & Kyle, 2021).[93]

However, a caveat about this is that, as noted in the paper, MTLD calculates partial factors for lexical remainders of a text (i.e., the words that do not form a full factor). These approximated remainders have a greater influence on MTLD the shorter the text is (i.e., the fewer full factors there are), which substantially increases the uncertainty associated with calculating MTLD on short texts, particularly when they are composed of only a remainder (McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020).

Initial estimates suggested that this measure should therefore be used primarily on texts longer than 100 words, but comprehensive recent research by Zenker and Kyle (2021) examined the use of MTLD in texts as short as 50 words, and found that it is fairly robust there too, though not perfectly so. This suggests that it can be reasonable to use MTLD on texts with 50 and potentially even fewer tokens, though this should be done with caution, and only for certain types of analyses, such as large-scale analyses of texts at the group level, as in the present study.[94]

In the present sample, some texts were shorter than 50 words, especially at the early proficiency levels, as shown in Figure 22 and Table 27. It is not possible to avoid this issue in the present sample, since selecting only texts above 50 words in advance (i.e., during the initial sample selection) would lead to a biased sample. Accordingly, it is important to replicate the present findings on additional learner samples, and preferably on samples that contain longer texts in particular.

Nevertheless, there are some things that are important to mention in this regard, as they strongly suggest—particularly when taken together—that the inclusion of short texts in the present sample does *not* invalidate the present findings:

[93] There are also some variations of MTLD that are relatively robust to text length, such as MTLD-MA-Wrap, but the research by Zenker and Kyle (2021) suggests that the original MTLD measure is more robust on short texts.

[94] Another situation where it might be reasonable to use MTLD on short texts is when it is used as one of only many different factors for assessing L2 proficiency.

1. Some of the issues associated with short texts were mitigated through the removal of outliers, as shown in the section on MTLD outliers.
2. In the study that found that MTLD is stable at the 50-word level (Zenker & Kyle, 2021), the 50-word threshold was the minimal number of words that was examined, which indicates that MTLD may also be stable at lower text lengths. Furthermore, as the analyses of Zenker and Kyle (2021) show, any potential slant in MTLD would be milder the closer the text is to the stabilization point in terms of number of words. Given this, and given that the texts in the sample were at least 20 words long, and their mean length was close to or above 50 at all proficiency levels (as shown in the table and figure below), there is a limit on any potential slant that could occur due to the use of shorter texts. As such, it is unlikely that any potential bias in the MTLD of the texts was substantial enough to change the main findings of the study, especially given that the findings regarding the association between L2 proficiency and MTLD were robust in the present sample and aligned with those of prior studies (Hout & Vermeer, 2010; Jarvis, 2013; Kojima & Yamashita, 2014; McCarthy & Jarvis, 2010; Murakami & Alexopoulou, 2016; Treffers-Daller et al., 2018; Yan et al., 2020).
3. Prior studies on the EFCAMDAT show that the MTLD of texts in it, including of short texts at the earliest proficiency levels, is strongly correlated with measures of syntactic complexity such as mean length clause and subordinate clause per T-unit, and that all these measures in return are correlated with proficiency (Alexopoulou et al., 2017; Murakami, 2014). This further suggests that MLTD scores on the short texts in EFCAMDAT do not distort the relation between lexical diversity and proficiency.
4. Generally, the lower learners’ L2 proficiency, the smaller their functional vocabulary, and the higher the rate of repetition in their writing. This means that, assuming the text length is held constant, as learners’ L2 proficiency *decreases*, we would expect the number of MTLD factors in their productions to increase, and therefore also expect the MTLD remainder to account for a smaller proportion of their total MTLD, which will lead to more accurate assessment of MTLD in short texts. This is particularly important in the context of the present analyses, since all the learners in the sample were at the beginner to intermediate range of L2 proficiency (CEFR A1–B2), so we would expect this to apply to them. Furthermore, as shown below, the shortest texts were generally written by the learners with the lowest L2 proficiency, where we would expect the highest rate of repetition, which reduces this issue when calculating MTLD on those texts.

5. Crucially, the main finding of the study, which was the lack of effect of lexical distance on MTLD, was consistent across all the CEFR levels, as shown in the Results section of the main paper (in the Figure with the scatterplots and linear models). This means that this finding held even in the higher-level sub-samples, where almost all texts are much longer than 50 words, and also at the highest levels, where nearly all texts are longer than 100 words. At the very least, this shows that this effect does not exist at those higher levels.
6. We also built supporting models, which included only texts from the A2–B2 CEFR range (Table 28 and Figures 23 and 24) and from the B1–B2 range (Table 29 and Figures 25 and 26). As shown in the Table below, these models contained substantially fewer texts below 50 words than the full models (which contained texts at the A1–B2 range), and in the case of the B1–B2 models there were almost no texts below 50 words, and most texts were longer than 100 words. Both sets of models replicated the main finding, and namely the lack of effect of lexical distance on lexical diversity.[95]
7. We also replicated the key analyses using MATTR as a measure of lexical diversity instead of MTLD, since it is also considered to be robust to short texts and variations in text length, and is calculated in a different way than MTLD (see §7.4.4.4 for these analyses, an explanation of MATTR, and the rationale behind focusing on MTLD in the main analyses). We found that there is a very high correlation between it and MTLD in the sample, and that the results are practically the same regardless of which of these measures of lexical diversity is used.

Overall, while we hope that the information that we present here will inform the broader discussion on lexical diversity, it is important to emphasize that we do *not* claim that it is always reasonable to use MTLD when analysing short texts. Rather, we acknowledge the substantial issues with calculating MTLD—as well as lexical diversity in general—on short texts. This means that it would be ideal to conduct any analyses pertaining to lexical diversity on texts that are at least 50 words longs—and preferably even 100 or more—which some of the texts in the present study are not. As such, it is recommended to replicate these findings on other learner samples, which contain longer texts.

[95] We do not attempt to assess the influence of L2 proficiency here, given the more limited range of L2 proficiency levels, and given the plateau in the association between L2 proficiency and lexical diversity that was found in the full analyses.

Nevertheless, when the evidence outlined above is taken in aggregate, it strongly suggests that any potential issues associated with short texts do *not* invalidate the findings of the present study. This evidence includes, most notably, the robustness of MTLD in short texts as indicated by Zenker and Kyle (2021), the association between lexical diversity and L2 proficiency (as well as other measures of linguistic complexity) that was found here and in previous studies, the replication of our key findings in sub-samples containing texts that are almost all longer than 50 words (as well as in sub-samples containing texts that are longer than 100 words), and the replication of our key findings when using MATTR as the measure of lexical diversity (§7.4.4.4).

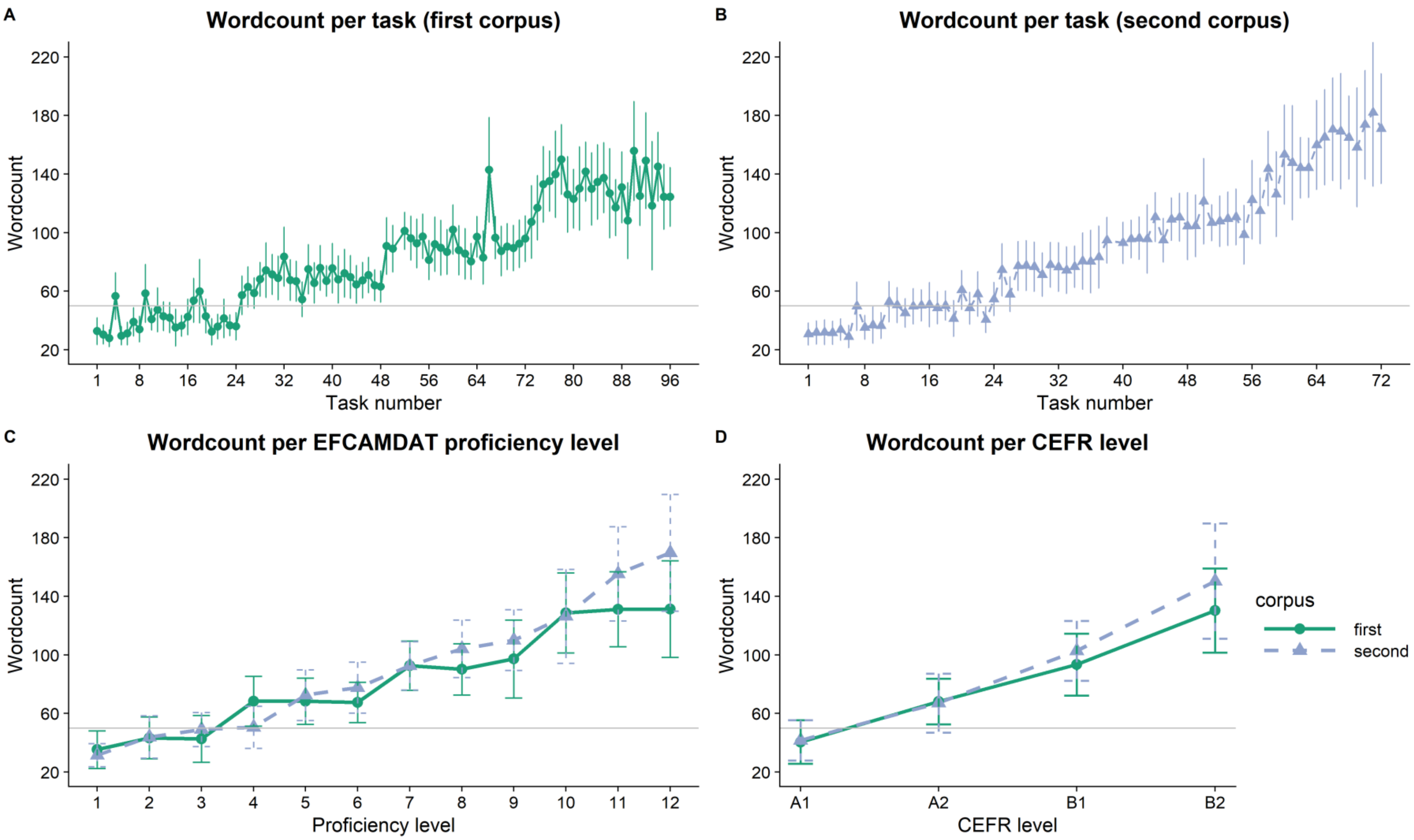


Figure 22. Mean wordcount of texts after removing outliers (error bars indicate one standard deviation). Listed per *task* in (A) and (B), per *EFCAMDAT proficiency level* in (C), and per *CEFR level* in (D). There are 8 tasks per EFCAMDAT proficiency level in the first corpus and 6 tasks per level in the second. There are 3 EFCAMDAT levels per CEFR level in both corpora. The grey line in each panel shows denotes a wordcount of 50.

Table 27. Statistics on the wordcounts of texts per CEFR (L2 proficiency) level in each corpus. *SD* denotes standard deviation, *Mdn* denotes median, *IQR* denotes the interquartile range, *Rng* denotes the full range, n_{total} denotes the total number of texts at that CEFR level, $n_{<50}$ denotes the number of texts at that level that have fewer than 50 words, and $p_{<50}$ denotes the percent of texts at that level that have fewer than 50 words.

	First corpus								Second corpus							
CEFR	Mean	SD	Mdn	IQR	Rng	n_{total}	$n_{<50}$	$p_{<50}$	Mean	SD	Mdn	IQR	Rng	n_{total}	$n_{<50}$	$p_{<50}$
A1	40.47	14.84	38	30-46	20-111	1921	1530	79.65	41.56	13.77	40	31-51	20-105	1508	1101	73.01
A2	68.11	15.54	68	58-75	22-133	2088	148	7.09	67.08	20.15	66	52-80	20-129	1553	334	21.51
B1	93.40	21.15	92	79-101	28-197	2042	13	0.64	102.80	20.40	104	91-115	31-188	1500	19	1.27
B2	130.35	28.76	130	109-149	21-245	2030	5	0.25	150.40	39.32	153	122-177	31-287	1568	4	0.26

Table 28. Results of the models that were calculated on the texts at the A2–B2 CEFR range.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	65.65	1.82	62.09 – 69.21	<0.001	0.00	-0.12 – 0.13	71.59	1.93	67.81 – 75.37	<0.001	0.00	-0.14 – 0.15
Lexical_distance	0.66	11.89	-22.64 – 23.96	0.956	-0.00	-0.07 – 0.06	5.87	13.43	-20.46 – 32.20	0.662	0.02	-0.07 – 0.10
L2_proficiency	3.57	0.54	2.50 – 4.64	<0.001	0.37	0.26 – 0.48	1.17	0.57	0.06 – 2.28	0.039	0.13	0.01 – 0.25
Lexical_distance * L2_proficiency	-1.40	1.38	-4.11 – 1.31	0.312	-0.01	-0.03 – 0.01	-0.61	1.65	-3.85 – 2.64	0.714	-0.01	-0.03 – 0.02
Random Effects												
σ^2	359.60						348.44					
τ_{00}	40.04_{Learner}						80.21_{Learner}					
	136.42_{Task}						109.29_{Task}					
	6.20_{L1}						7.75_{L1}					
ICC	0.34						0.36					
N	9_{L1}						9_{L1}					
	3882_{Learner}						3182_{Learner}					
	71_{Task}						53_{Task}					
Observations	6160						4621					
Mar. R^2 / Cond. R^2	0.135 / 0.427						0.017 / 0.372					

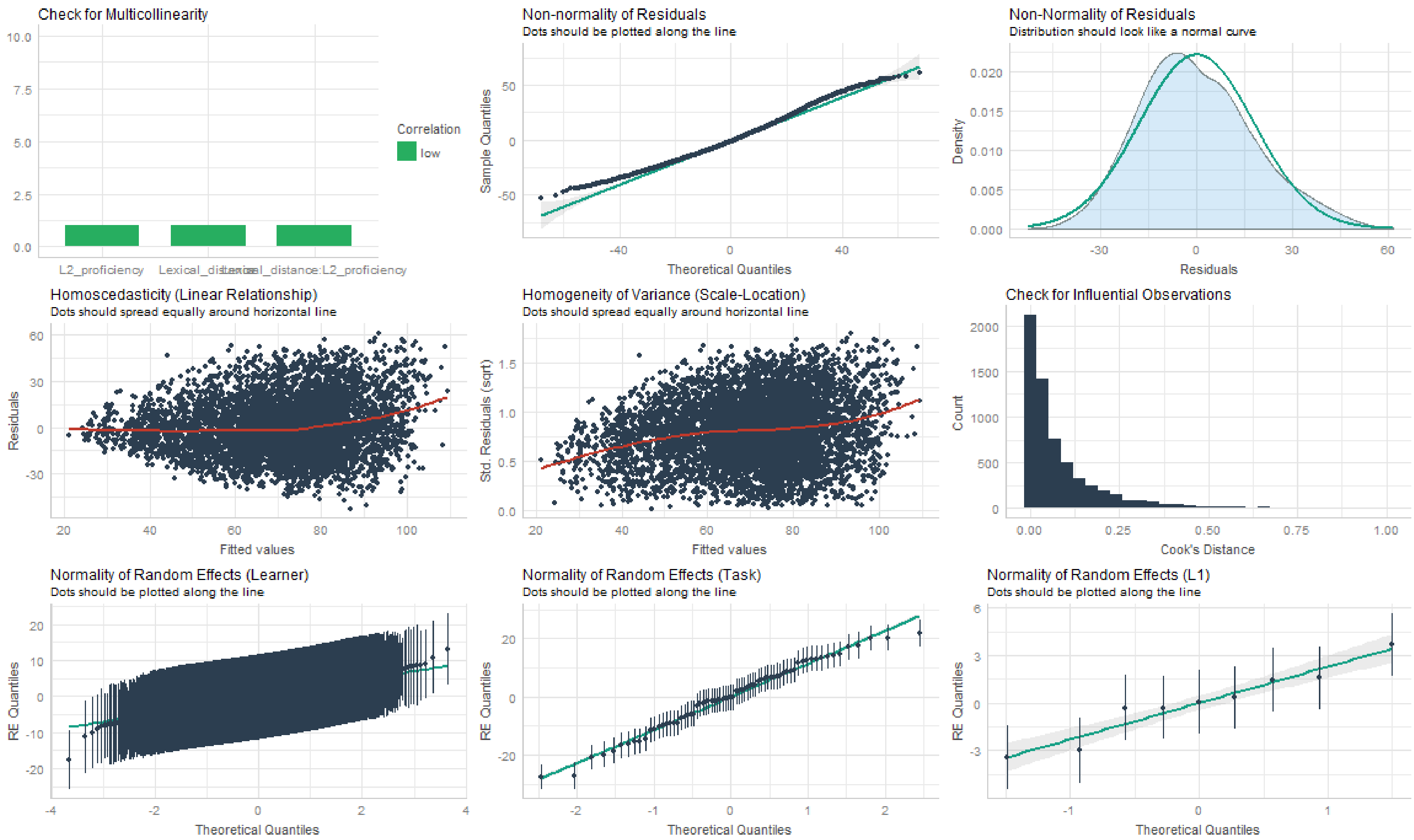


Figure 23. Assumption checks for the model with texts at the A2–B2 range, in the first corpus.

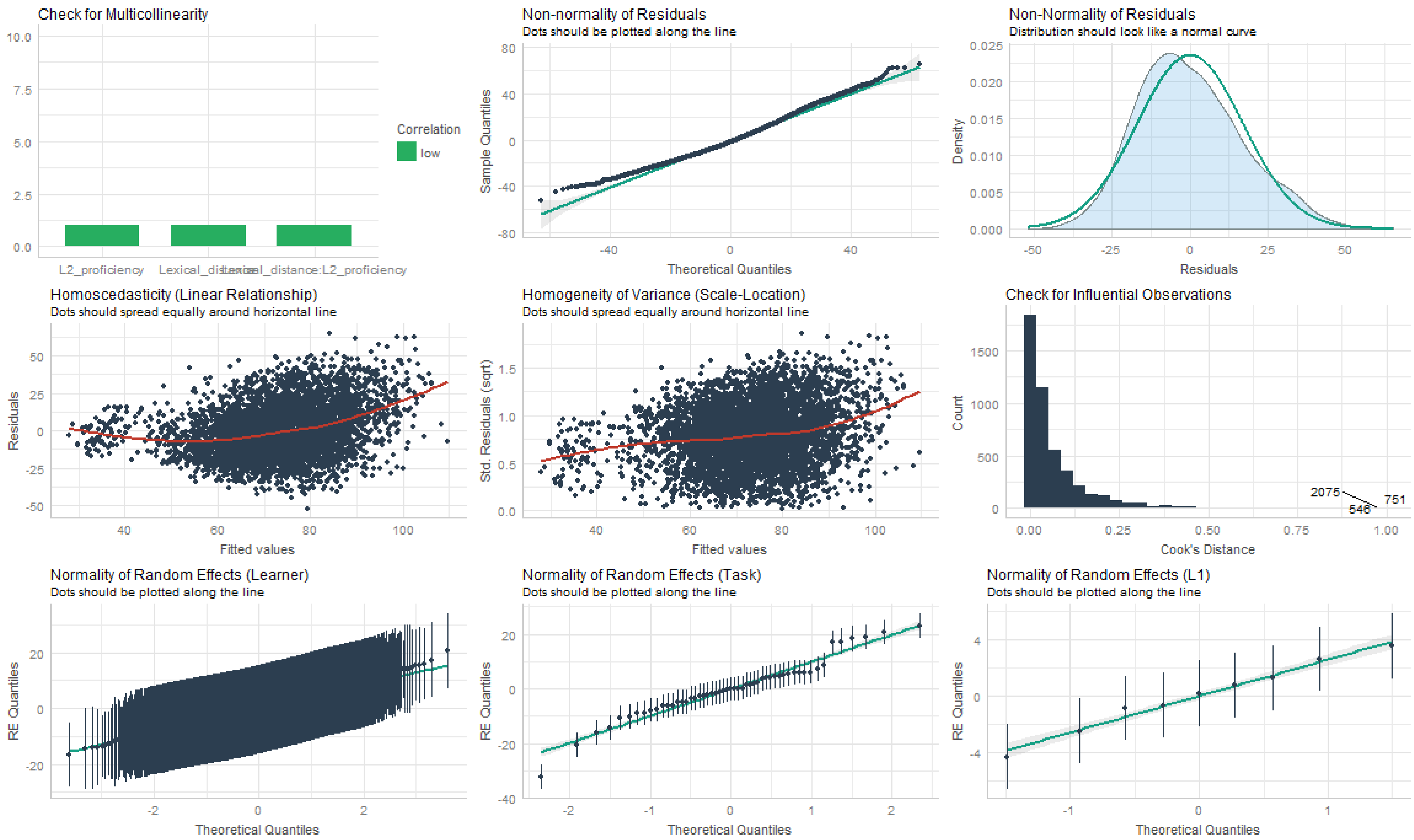


Figure 24. Assumption checks for the model with texts at the A2–B2 range, in the second corpus.

Table 29. Results of the models that were calculated on the texts at the B1–B2 CEFR range.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	71.77	3.28	65.34 – 78.20	<0.001	0.01	-0.15 – 0.16	75.51	2.81	69.99 – 81.02	<0.001	0.00	-0.14 – 0.15
Lexical_distance	6.83	17.07	-26.62 – 40.28	0.689	-0.01	-0.10 – 0.08	5.48	16.92	-27.68 – 38.65	0.746	0.01	-0.08 – 0.11
L2_proficiency	1.76	0.90	-0.01 – 3.54	0.051	0.13	-0.00 – 0.25	0.18	0.76	-1.32 – 1.68	0.812	0.01	-0.10 – 0.13
Lexical_distance * L2_proficiency	-3.23	2.74	-8.60 – 2.15	0.239	-0.02	-0.05 – 0.01	-0.44	3.11	-6.55 – 5.66	0.887	-0.00	-0.04 – 0.03
Random Effects												
σ^2	393.08						344.15					
τ_{00}	52.70 $_{\text{Learner}}$						91.71 $_{\text{Learner}}$					
	103.60 $_{\text{Task}}$						52.20 $_{\text{Task}}$					
	10.23 $_{\text{L1}}$						8.66 $_{\text{L1}}$					
ICC	0.30						0.31					
N	9 $_{\text{L1}}$						9 $_{\text{L1}}$					
	2335 $_{\text{Learner}}$						1978 $_{\text{Learner}}$					
	47 $_{\text{Task}}$						35 $_{\text{Task}}$					
Observations	4072						3068					
Mar. R^2 / Cond. R^2	0.016 / 0.309						0.000 / 0.307					

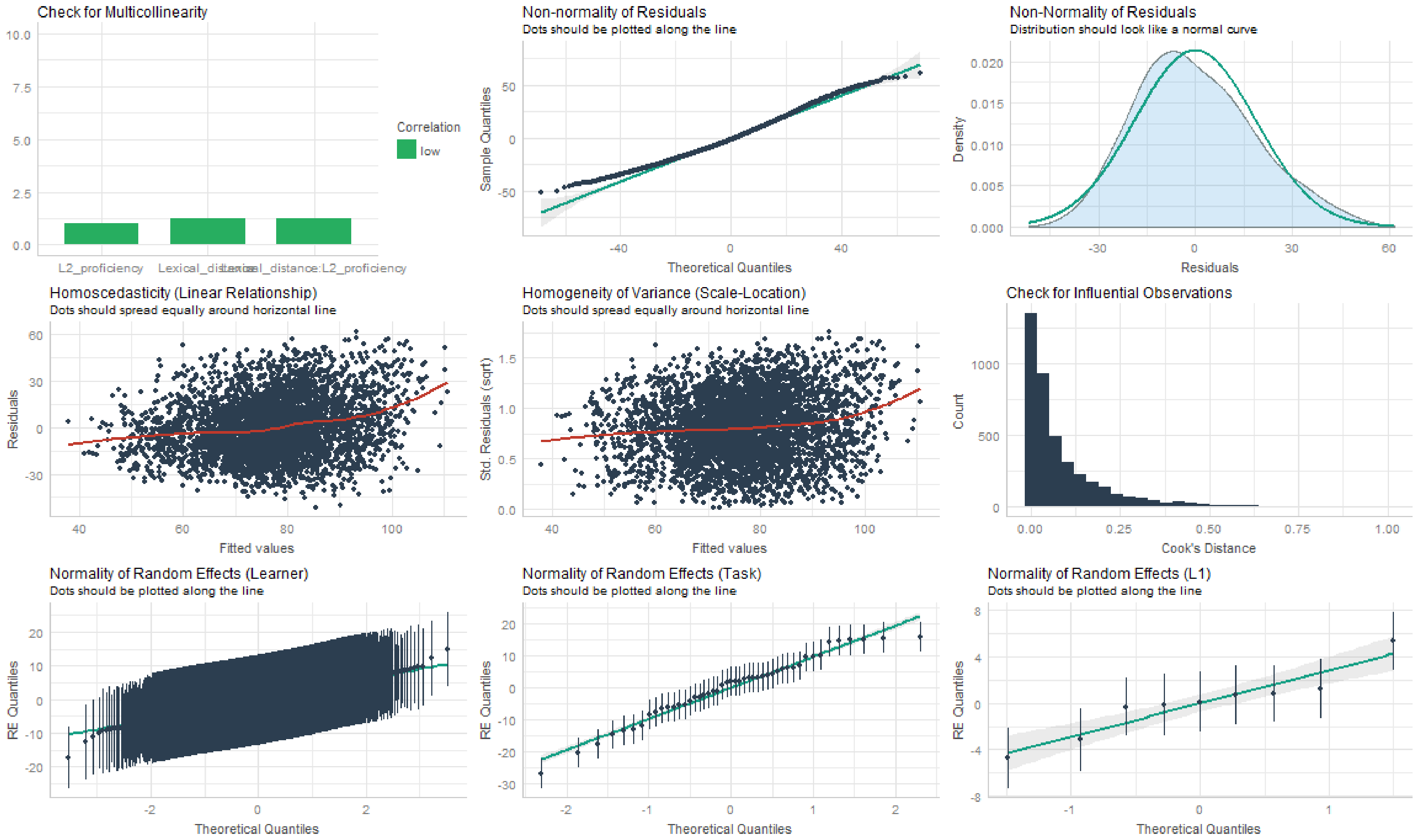


Figure 25. Assumption checks for the model with texts at the B1–B2 range, in the first corpus.

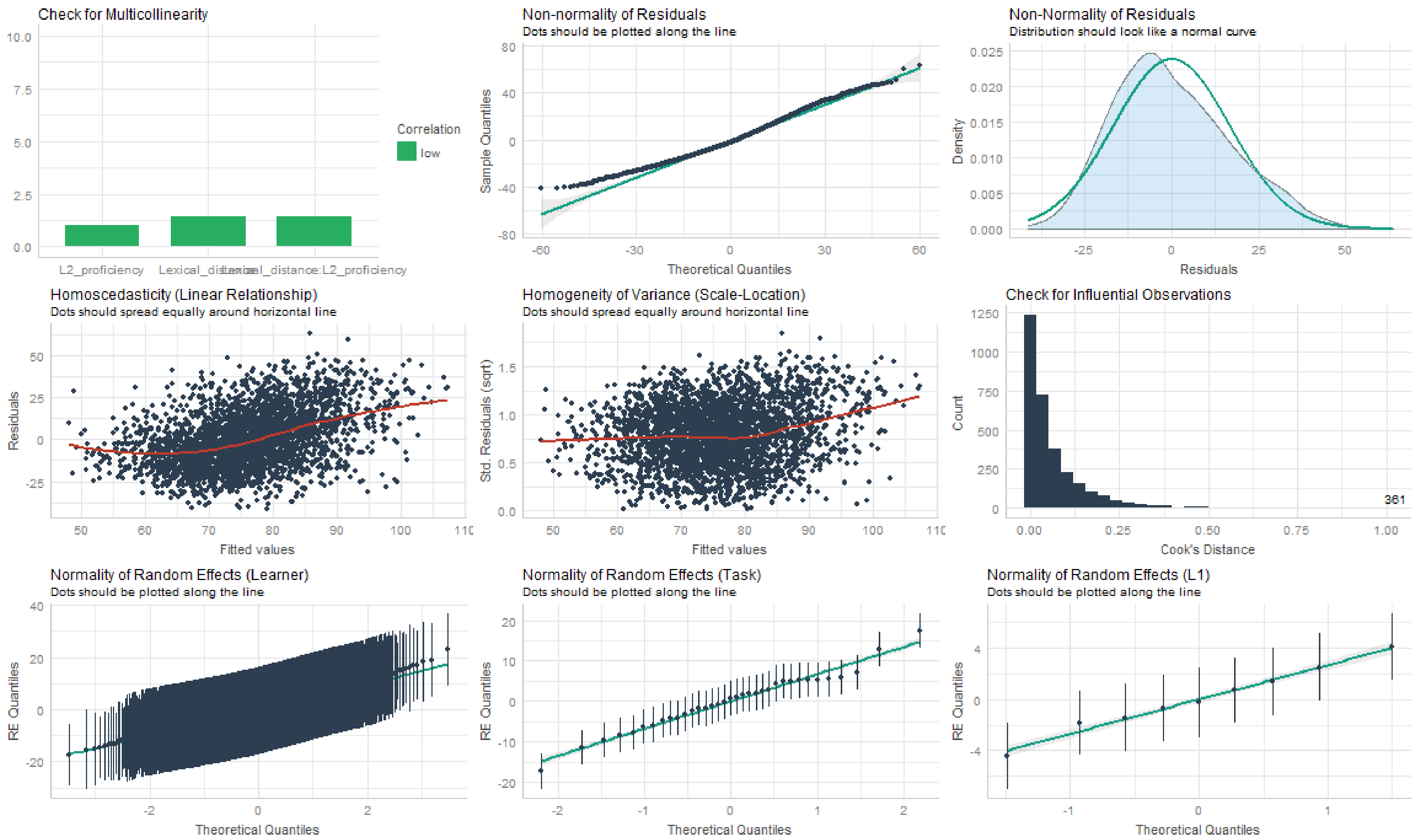


Figure 26. Assumption checks for the model with texts at the B1–B2 range, in the second corpus.

7.4.4 Supporting models

This section contains supporting mixed-effects models that were used to complement the main models in the paper. For all models, the following holds unless noted otherwise:

- Under fixed effects, *lexical_distance* is the mean lexical distance between the L1 and English (0–1), and *L2_proficiency* is the EFCAMDAT level associated with each text (1–12).
- In the statistics, *std. B* and *std. 95% CI* provide information on the standardized coefficients, which were calculated by refitting the model on standardized data.
- Under random effects, σ^2 denotes the residual variance, τ_{00} denotes between-subjects (or groups) variance, τ_{11} denotes the random-slope variance, ρ_{01} denotes the random-slope-intercept correlation, *ICC* denotes the intraclass correlation coefficient, and *N* denotes the number of data points within each sampling unit.
- Finally, *observations* denotes the total number of texts in each sample, *Mar. [Marginal]* R^2 denotes the proportion of the variance described by the fixed effects, and *Cond. [Conditional]* R^2 denotes the proportion of the variance described by both the fixed and random effects.

Interpretation of the associated assumption checks is based on the approach outlined in §7.4.1 of this document.

7.4.4.1 Random slopes models

Several mixed-effect models with random slopes were considered and ruled out (Matuschek et al., 2017; Winter, 2019).

First, in the case of random slopes of *L2 proficiency* for *learner*, the number of observations was smaller than the number of parameters pertaining to random effects, since most learners only have a single text in the sample, as shown in the “Sample information” document in the study’s *Open Science Framework* (OSF) repository, so the models did not converge, and had an error indicating that “the random-effects parameters and the residual variance (or scale parameter) are probably unidentifiable”.

Second, in the case of random slopes of *lexical distance* for *task*, the models did converge, but a comparison showed that the models with only random intercepts were preferable. Specifically, we compared the AIC and BIC across models, as shown in Table 30.

Table 30. Comparisons of AIC and BIC across models. Both measures were used, as suggested in Kuha (2004). Data was generated in R using the *AIC* and *BIC* functions.

Corpus	Model	df	AIC	Δ AIC	BIC	Δ BIC
First	Only intercept	8	71071.51	0.45	71127.49	-
First	With slope	10	71071.06	-	71141.03	13.54
Second	Only intercept	8	53972.68	-	54026.45	-
Second	With slope	10	53973.58	0.90	54040.79	14.34

Note. ΔAIC is calculated by subtracting the AIC of a given model from the AIC of the model with the minimal AIC in that corpus. Accordingly, no ΔAIC is listed for the model with the minimal AIC in a corpus. The same is the case for ΔBIC.

The difference in AIC was not substantial in either corpus (Δ < 2).[96] However, the difference in BIC presented strong evidence (ΔBIC > 10) in favor of the intercepts-only model, in both corpora. Given this, the more parsimonious intercepts-only models were used as the main ones for the study. However, this choice does not substantially affect the findings of the study, as shown in Table 31, which contains the results of these random-slopes models (with the corresponding assumption checks in Figures 27 and 28), which mirror the results of the main models, when it comes to the non-significant and negligible effect of lexical distance and of its interaction with L2 proficiency, as well as the significant and substantial effects of L2 proficiency and task. This is unsurprising, since the main concern with not including potential random slopes is an increased rate of Type I error (Matuschek et al., 2017; Winter, 2019), but the effects of lexical distance and its interaction were non-significant, while the findings for L2 proficiency were robust in terms of significance.

[96] Interpretations of the differences in AIC/BIC are based on Fabozzi et al. (2014).

Table 31. Results of the models with random slopes of *lexical distance* for *task*.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	64.91	1.43	62.10 – 67.71	<0.001	0.01	-0.10 – 0.12	66.70	1.58	63.61 – 69.80	<0.001	0.00	-0.12 – 0.13
Lexical_distance	-2.52	10.37	-22.84 – 17.79	0.808	-0.01	-0.06 – 0.05	4.26	11.71	-18.69 – 27.21	0.716	0.01	-0.06 – 0.08
L2_proficiency	3.82	0.36	3.12 – 4.52	<0.001	0.50	0.41 – 0.59	3.10	0.39	2.34 – 3.86	<0.001	0.43	0.32 – 0.54
Lexical_distance * L2_proficiency	-0.11	0.99	-2.05 – 1.84	0.914	-0.00	-0.02 – 0.02	0.17	1.11	-2.01 – 2.35	0.879	0.00	-0.02 – 0.02
Random Effects												
σ^2	336.54						312.14					
τ_{00}	34.03 $_{\text{Learner}}$						66.37 $_{\text{Learner}}$					
	139.77 $_{\text{Task}}$						124.06 $_{\text{Task}}$					
	4.75 $_{\text{L1}}$						6.05 $_{\text{L1}}$					
τ_{11}	247.89 $_{\text{Task.Lexical\_distance}}$						179.03 $_{\text{Task.Lexical\_distance}}$					
ρ_{01}	-0.18 $_{\text{Task}}$						0.36 $_{\text{Task}}$					
ICC	0.35						0.39					
N	9 $_{\text{L1}}$						9 $_{\text{L1}}$					
	5385 $_{\text{Learner}}$						4357 $_{\text{Learner}}$					
	95 $_{\text{Task}}$						71 $_{\text{Task}}$					
Observations	8081						6129					
Mar. R^2 / Cond. R^2	0.249 / 0.511						0.184 / 0.500					

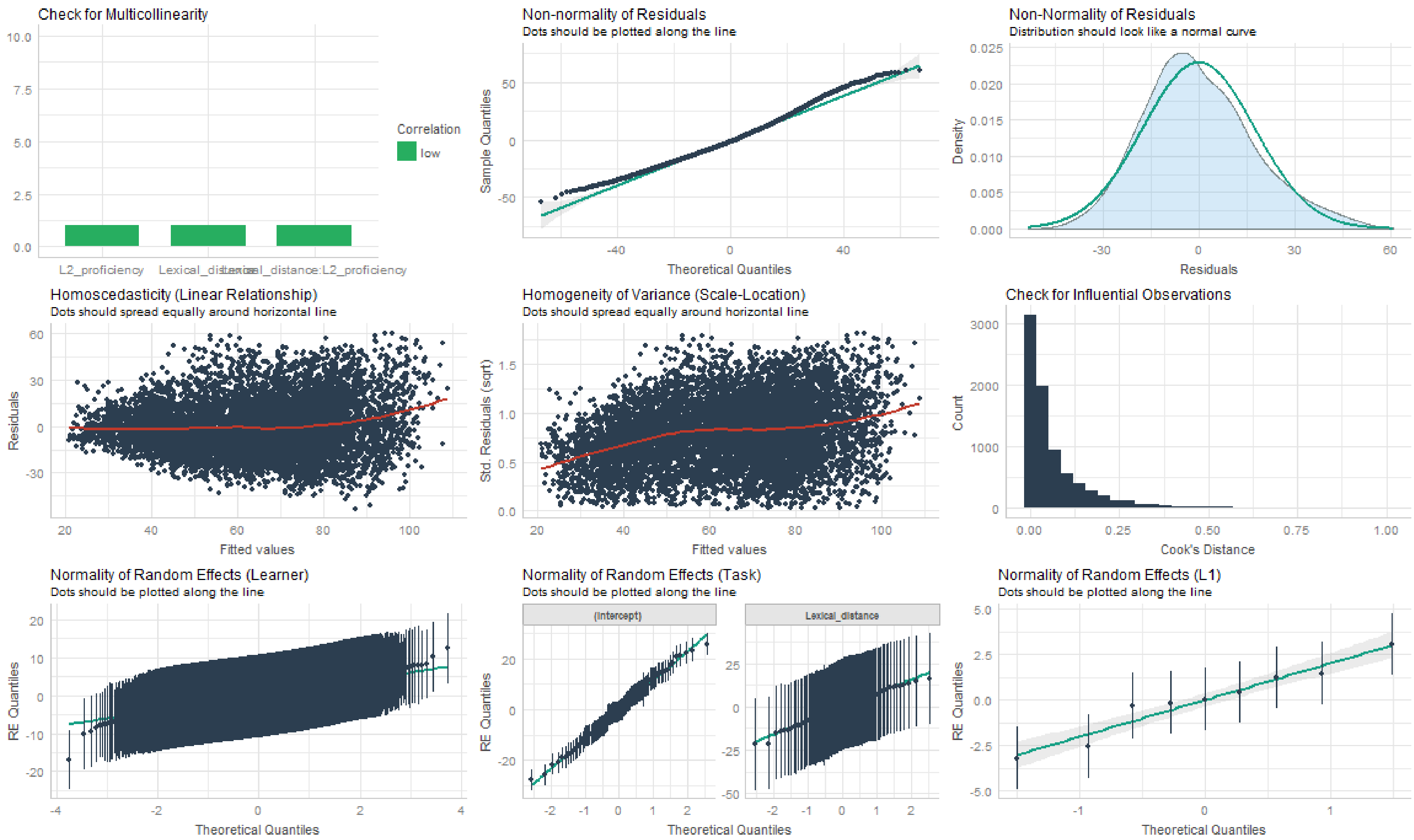


Figure 27. Assumption checks for the models with random slopes of *lexical distance* for *task* in the first corpus.

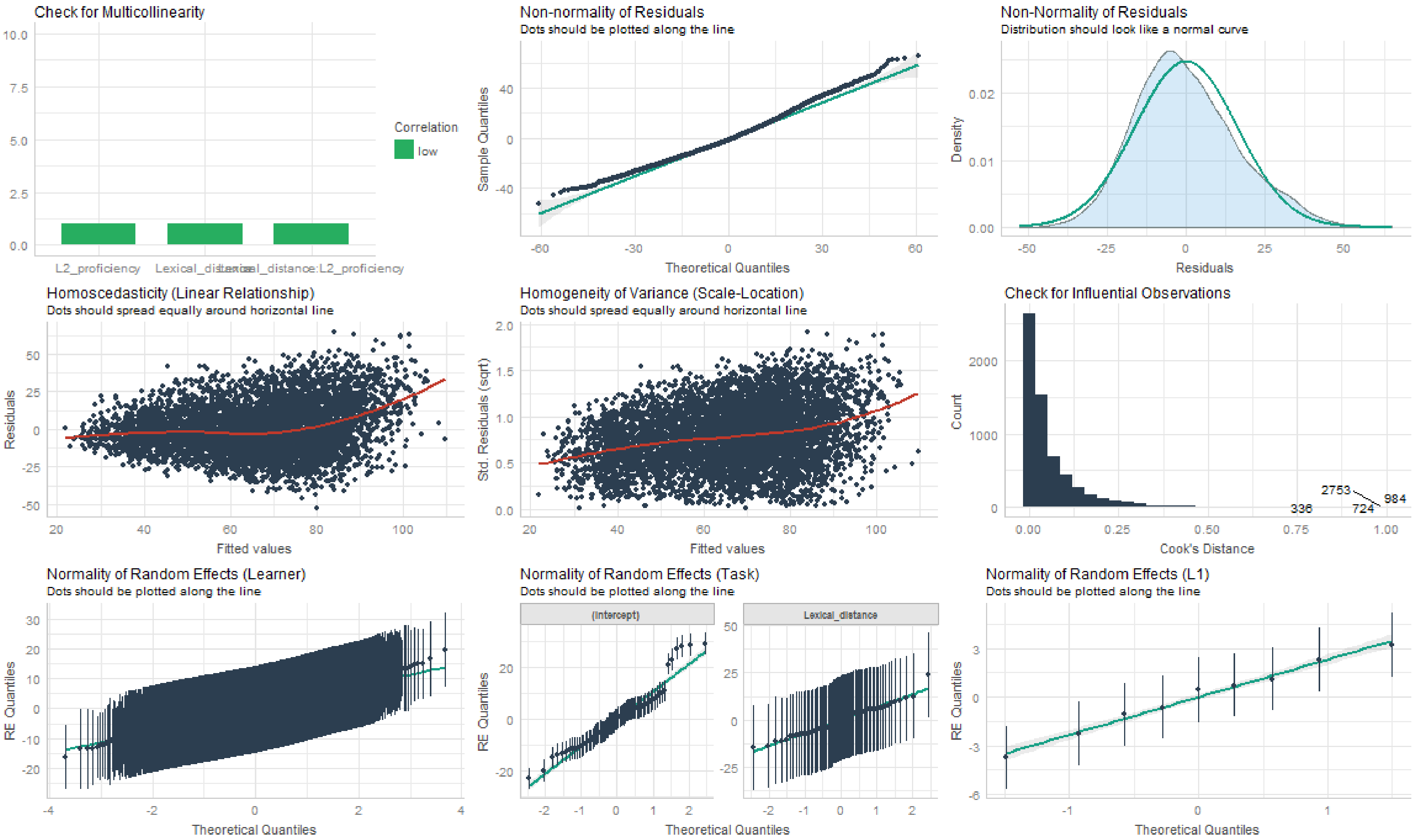


Figure 28. Assumption checks for the models with random slopes of *lexical distance* for *task* in the second corpus.

Finally, in the case of random slopes of *L2 proficiency* for *L1*, the models had singularity issues in the both the first and second corpora, indicating that "the parameters are on the boundary of the feasible parameter space: variances of one or more linear combinations of effects are (close to) zero". This suggests that inferences that are drawn from these models may not be reliable, so we did also not include these random slopes in our main analyses. Nevertheless, as in the case of random slopes of *lexical distance* for *task*, the results of these models closely mirror those of the main models, as shown in Table 32 (the corresponding assumption checks appear in Figures 29 and 30).

Overall, the information in this section provides support for the use of the random-intercepts model that appeared in the paper. Furthermore, it suggests that including potential random slopes does not substantially change the findings of the study. Nevertheless, as recommended by Meteyard & Davies (2020), we acknowledge that the choice to use this particular model specification represents one justified path out of several possible ones.

Table 32. Results of the models with random slopes of *L2 proficiency* for *L1*.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	64.87	1.45	62.03 – 67.71	<0.001	0.01	-0.10 – 0.12	66.70	1.59	63.59 – 69.81	<0.001	0.00	-0.12 – 0.13
Lexical_distance	-2.59	10.66	-23.48 – 18.29	0.808	-0.01	-0.07 – 0.05	4.40	11.79	-18.70 – 27.51	0.709	0.01	-0.06 – 0.08
L2_proficiency	3.80	0.37	3.08 – 4.53	<0.001	0.50	0.40 – 0.59	3.10	0.39	2.33 – 3.87	<0.001	0.43	0.32 – 0.54
Lexical_distance * L2_proficiency	-0.21	1.60	-3.34 – 2.92	0.894	-0.00	-0.03 – 0.03	0.21	1.26	-2.26 – 2.69	0.867	0.00	-0.02 – 0.03
Random Effects												
σ^2	338.07						312.64					
τ_{00}	32.99 $_{\text{Learner}}$						66.65 $_{\text{Learner}}$					
	139.82 $_{\text{Task}}$						124.09 $_{\text{Task}}$					
	5.18 $_{\text{L1}}$						6.26 $_{\text{L1}}$					
τ_{11}	0.09 $_{\text{L1.L2\_proficiency}}$						0.03 $_{\text{L1.L2\_proficiency}}$					
ρ_{01}	1.00 $_{\text{L1}}$						1.00 $_{\text{L1}}$					
N	9 $_{\text{L1}}$						9 $_{\text{L1}}$					
	5385 $_{\text{Learner}}$						4357 $_{\text{Learner}}$					
	95 $_{\text{Task}}$						71 $_{\text{Task}}$					
Observations	8081						6129					
Mar. R^2 / Cond. R^2	0.334 / NA						0.269 / NA					

Note. Some of the statistics are missing or equal to zero due to the singularity and convergence issues with the models, which are specified in the body of the text.

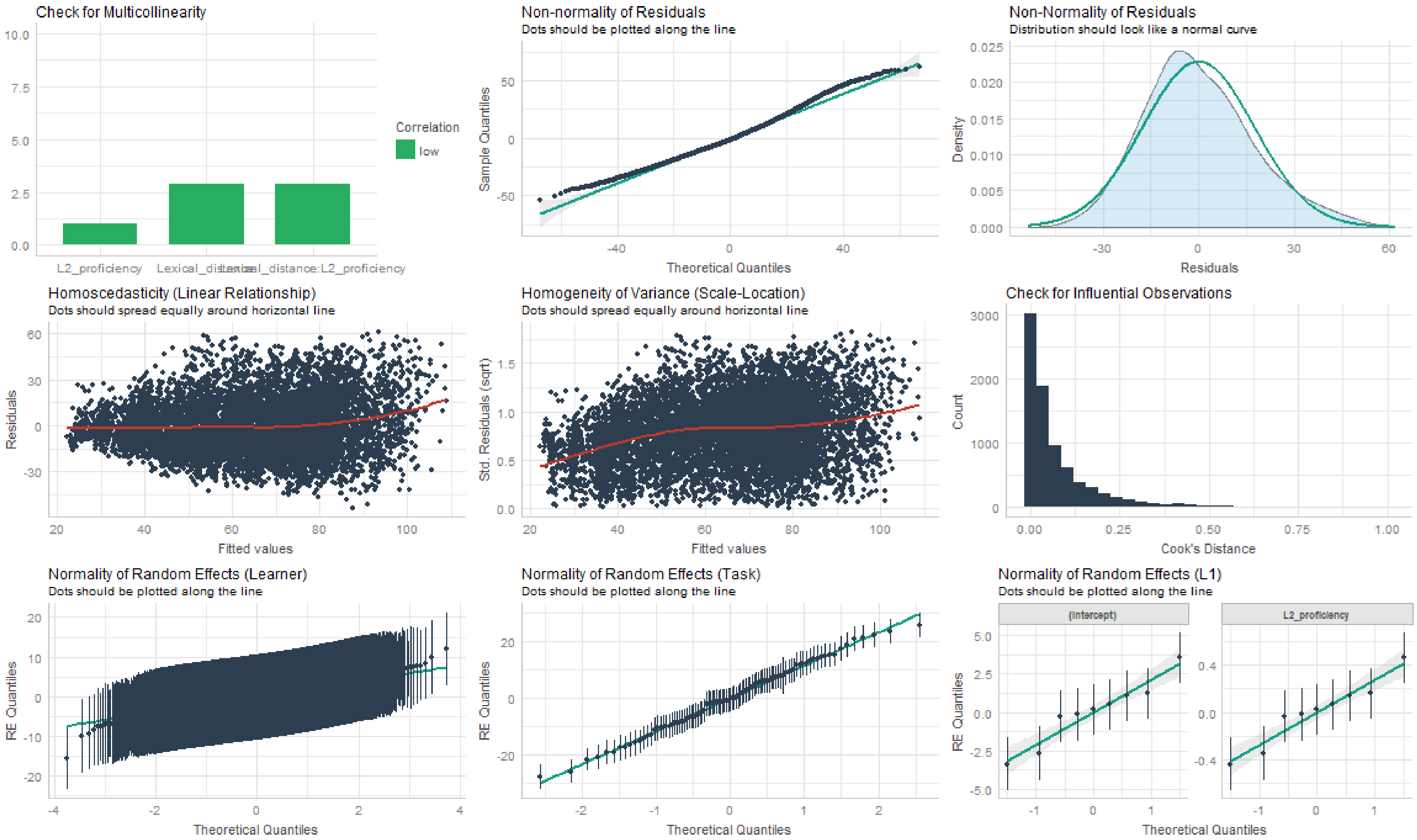


Figure 29. Assumption checks for the models with random slopes of *L2 proficiency* for *L1* in the first corpus.

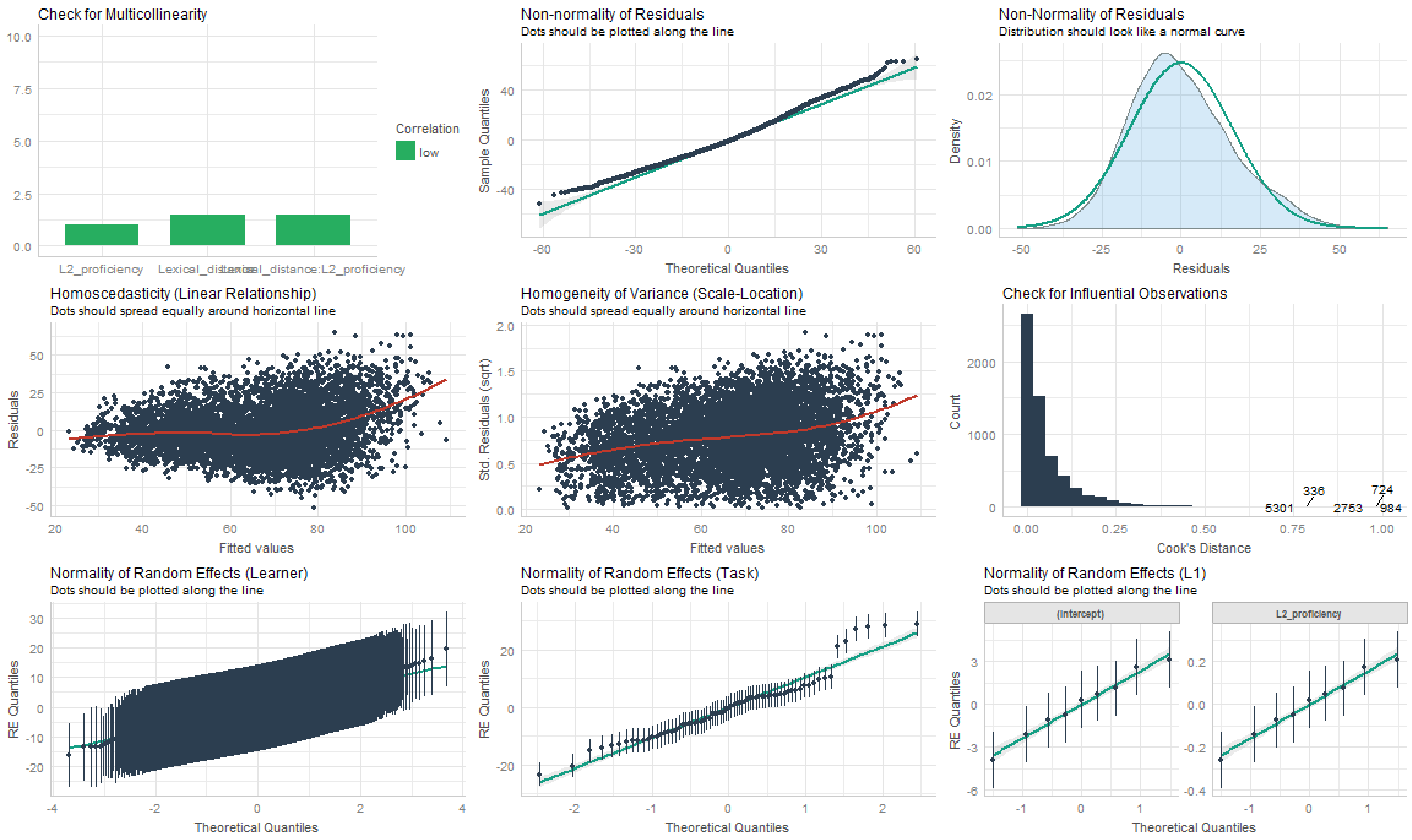


Figure 30. Assumption checks for the models with random slopes of *L2 proficiency* for *L1* in the first corpus.

7.4.4.2 Baseline models

The baseline models were models with no lexical distance as a predictor, and consequently also no interaction between lexical distance and L2 proficiency.[97] The goal of comparing the main models to these models was to provide further insights into the effects of lexical distance, by showing what happens when it is excluded from the model.

Table 33 shows the results of the baseline model. As expected, given the null effect of lexical distance and of its interaction with L2 proficiency in the main models, the results here closely mirror those of the main models. Figures 31 and 32 show the assumption checks for the models, which also closely mirror those of the main models (though there is no check for collinearity, since there is only one predictor). In addition, Table 34 contains the AIC/BIC comparison with the main models, based on the approach outlined in the previous section on the random-slopes models. The results of this comparison are mixed, with AIC providing weak support for the main models, and BIC providing strong support for the baseline models, so overall there was stronger support for the baseline models. This was expected given the lack of effect of lexical distance and of its interaction with L2 proficiency, and suggests that the weak support based on AIC is due to overfitting in the main models.

[97] Recall that, as noted in the main paper, this interaction term is meant to determine whether the effect of lexical distance varies as a function of L2 proficiency (i.e., whether the effect of lexical distance becomes weaker as L2 proficiency increases), since prior studies indicated that the expected L1 effects are generally stronger at lower proficiency levels.

Table 33. Results of the baseline models (with no lexical distance).

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	64.91	1.41	62.15 – 67.67	<0.001	0.01	-0.10 – 0.11	66.70	1.55	63.66 – 69.75	<0.001	0.00	-0.12 – 0.12
L2_proficiency	3.82	0.36	3.12 – 4.52	<0.001	0.50	0.41 – 0.59	3.10	0.39	2.34 – 3.86	<0.001	0.43	0.32 – 0.54
Random Effects												
σ^2	337.83						313.05					
τ_{00}	34.04_{Learner}						66.37_{Learner}					
	139.67_{Task}						124.02_{Task}					
	4.13_{L1}						5.33_{L1}					
ICC	0.34						0.38					
N	9_{L1}						9_{L1}					
	5385_{Learner}						4357_{Learner}					
	95_{Task}						71_{Task}					
Observations	8081						6129					
Mar. R^2 / Cond. R^2	0.249 / 0.508						0.184 / 0.498					

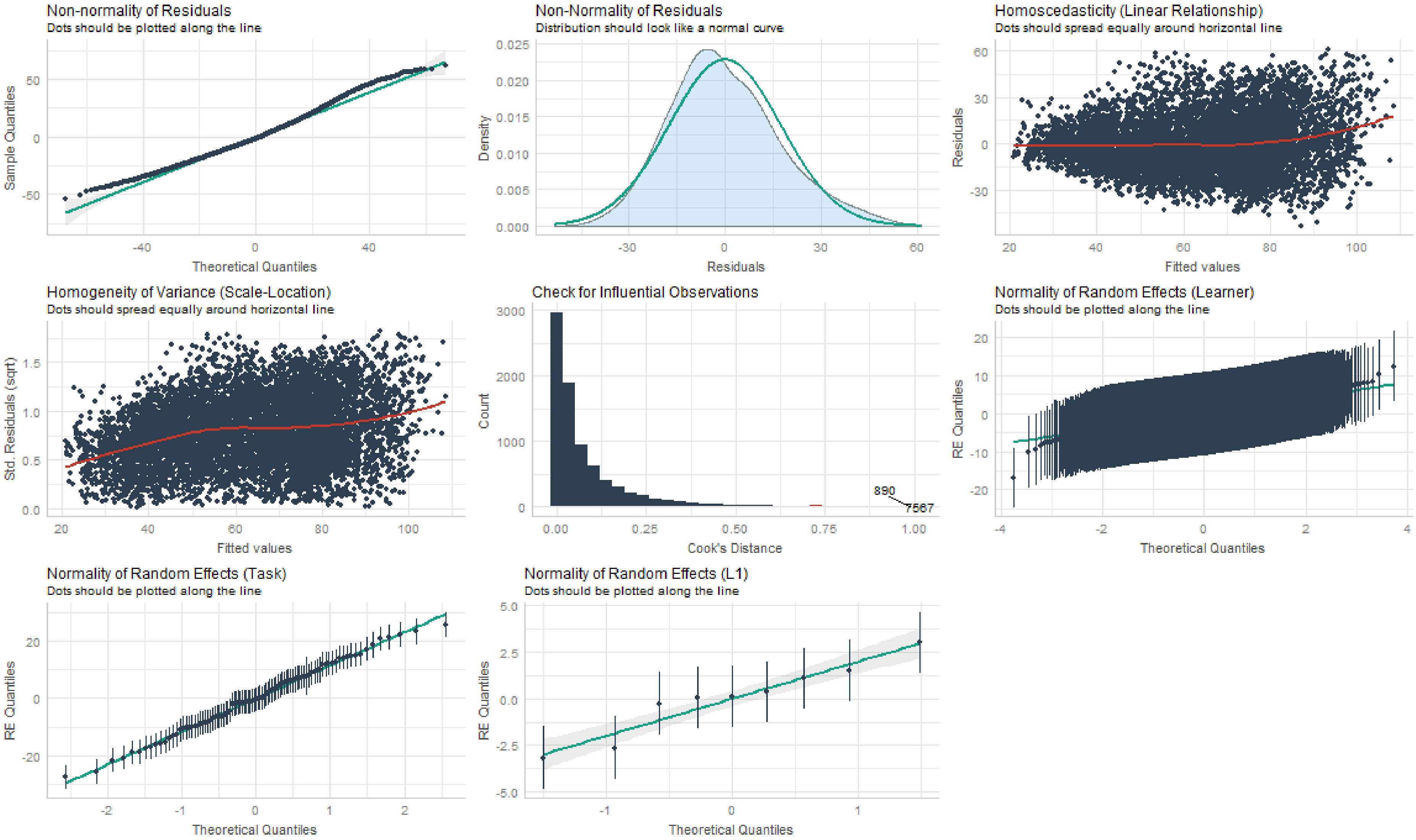


Figure 31. Assumption checks for the baseline models in the first corpus.

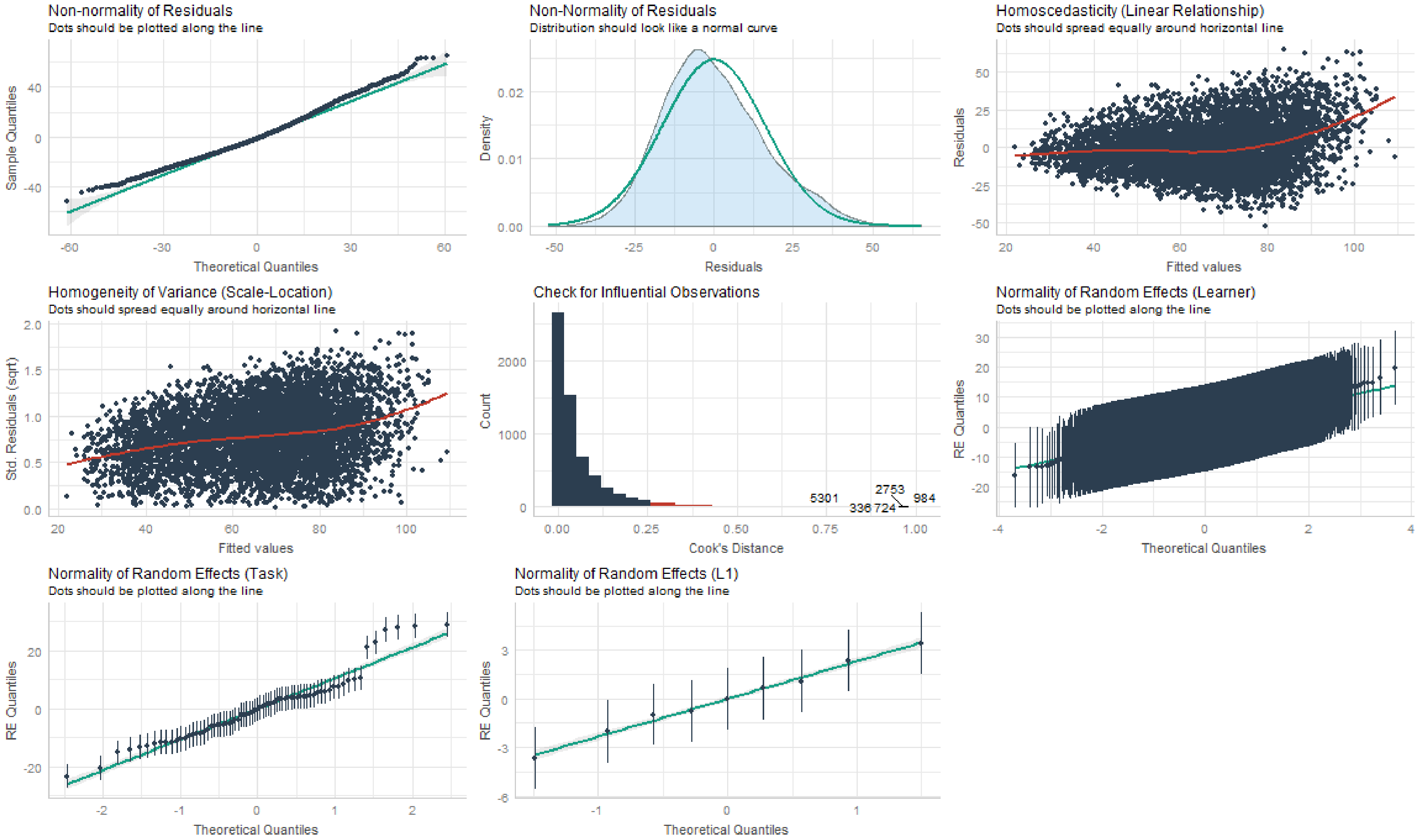


Figure 32. Assumption checks for the baseline models in the second corpus.

Table 34. Comparisons of AIC and BIC across models. Both measures were used, as suggested in Kuha (2004). Data was generated in R using the *AIC* and *BIC* functions.

Corpus	Model	df	AIC	Δ AIC	BIC	Δ BIC
First	Main models	8	71071.51	-	71127.49	9.92
First	Baseline	6	71075.59	4.07	71117.57	-
Second	Main models	8	53972.68	-	54026.45	8.71
Second	Baseline	6	53977.41	4.73	54017.74	-

Note. ΔAIC is calculated by subtracting the AIC of a given model from the AIC of the model with the minimal AIC in that corpus; accordingly, no ΔAIC is listed for the model with the minimal AIC in the corpus. The same is the case for ΔBIC.

7.4.4.3 Models with binary distance

In these models, rather than use phonological LDN based on Swadesh lists as the measure of lexical distance, we used a binary measure of lexical distance, based on whether the L1 was an Indo-European language like English or not. The Indo-European L1s included German, French, Italian, Portuguese, Spanish, and Russian, and were coded as *0* in the models. The non-Indo-European L1s included Arabic, Japanese, and Mandarin, and were coded as *1*. The goal of these models was to check if the findings hold when the L1s are categorized based on general linguistic relation to English.

Table 35 shows the results of the models that used this distance measure, and Figures 33 and 34 show the associated assumption checks. The results of these models closely mirror the results of the main models, as there are no substantial differences in any of the patterns. This was expected, given the robustness of the lack of effect in the main models and associated analyses (especially the estimated marginal means), and given that the binary distances align directly with the phonological LDN, in the sense that all the Indo-European L1s were lexically closer to English than all the non-Indo-European L1s.

Table 35. Results of the models that use binary distance as a measure of lexical distance, based on whether the L1 was an Indo-European language like English or not. The Indo-European L1s included German, French, Italian, Portuguese, Spanish, and Russian, and were coded as *0* in the models. The non-Indo-European L1s included Arabic, Japanese, and Mandarin, and were coded as *1*.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	65.25	1.52	62.28 – 68.22	<0.001	0.01	-0.10 – 0.12	66.32	1.68	63.03 – 69.61	<0.001	0.00	-0.12 – 0.13
Lexical_distance	-1.08	1.58	-4.18 – 2.02	0.494	-0.02	-0.07 – 0.04	1.14	1.80	-2.38 – 4.67	0.524	0.02	-0.04 – 0.09
L2_proficiency	3.88	0.36	3.18 – 4.58	<0.001	0.50	0.41 – 0.59	3.10	0.39	2.33 – 3.87	<0.001	0.43	0.32 – 0.54
Lexical_distance * L2_proficiency	-0.20	0.14	-0.48 – 0.08	0.159	-0.01	-0.03 – 0.00	-0.01	0.16	-0.33 – 0.31	0.966	-0.00	-0.02 – 0.02
Random Effects												
σ^2	337.78						313.04					
τ_{00}	34.05 $_{\text{Learner}}$						66.45 $_{\text{Learner}}$					
	139.71 $_{\text{Task}}$						124.01 $_{\text{Task}}$					
	4.52 $_{\text{L1}}$						5.80 $_{\text{L1}}$					
ICC	0.35						0.39					
N	9 $_{\text{L1}}$						9 $_{\text{L1}}$					
	5385 $_{\text{Learner}}$						4357 $_{\text{Learner}}$					
	95 $_{\text{Task}}$						71 $_{\text{Task}}$					
Observations	8081						6129					
Mar. R^2 / Cond. R^2	0.249 / 0.509						0.184 / 0.499					

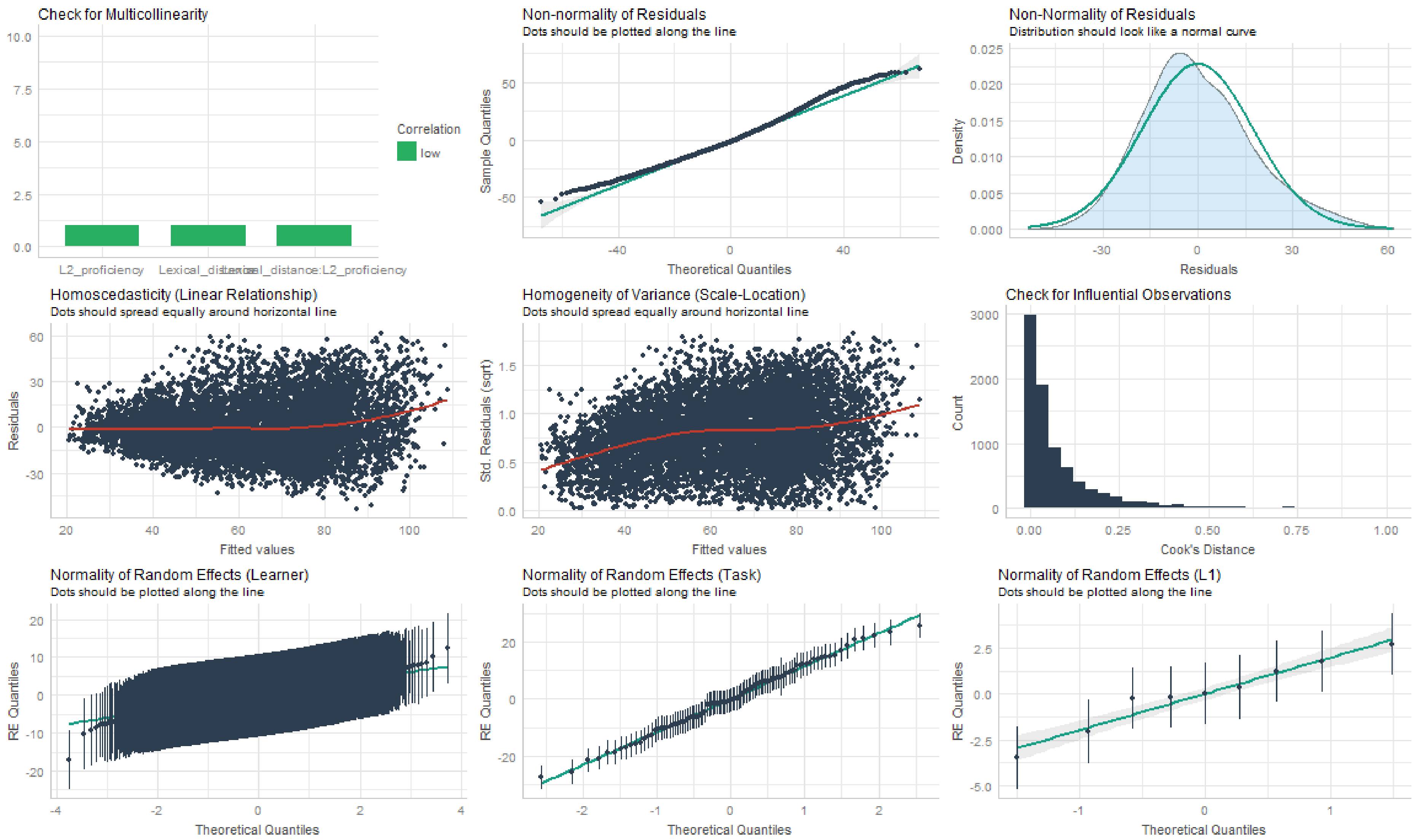


Figure 33. Assumption checks for the model with binary distance as a predictor, in the first corpus.

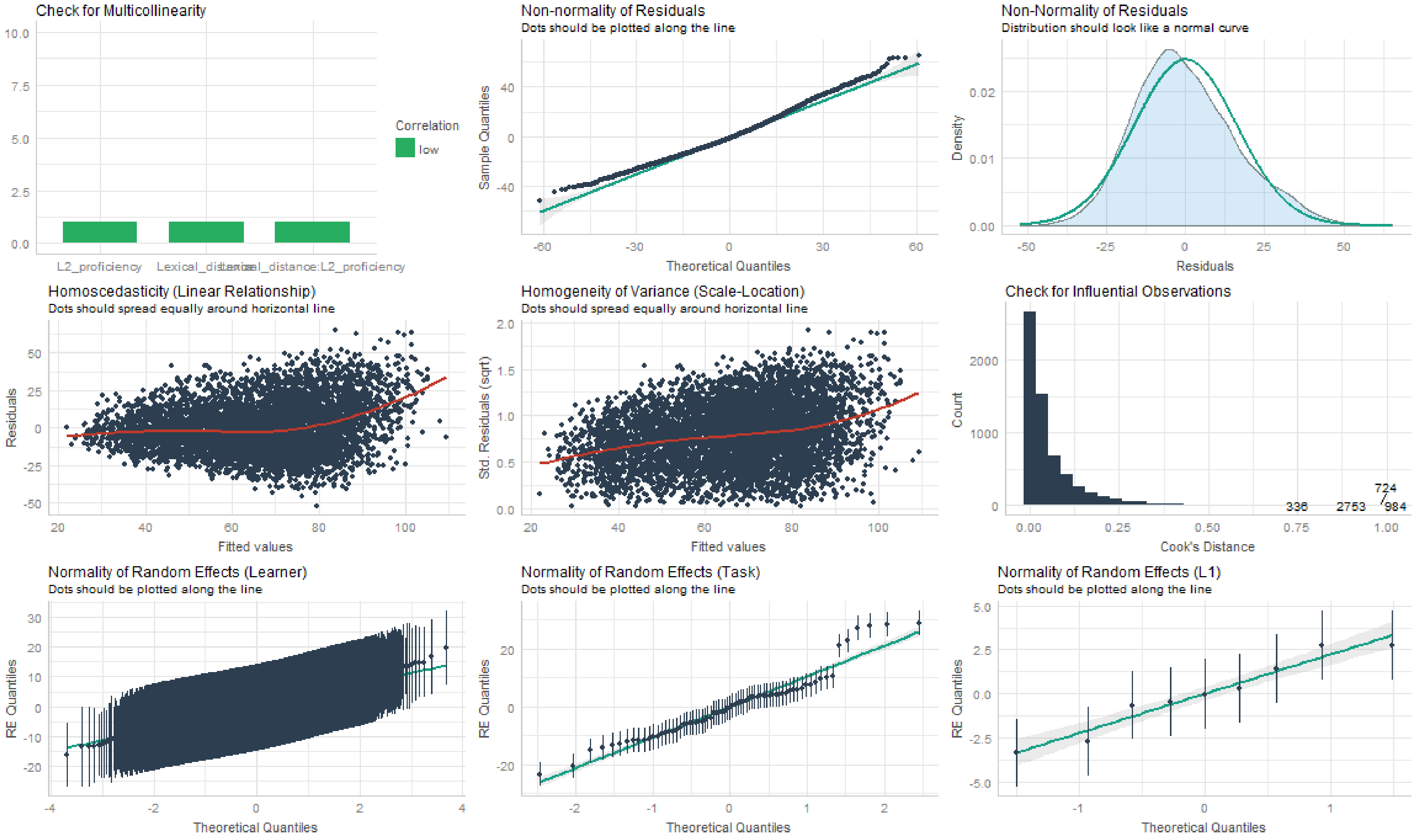


Figure 34. Assumption checks for the model with binary distance as a predictor, in the second corpus.

7.4.4.4 MATTR-based models

7.4.4.4.1 Rationale

In the main analyses in our study, we used *measure of textual lexical diversity* (MTLD) as a measure of lexical diversity, for two main reasons. First, there is substantial prior research on it, which facilitates the comparison of our findings with those of others (e.g., Treffers-Daller et al., 2018), and which shows that MTLD is strongly correlated with other common measures of lexical diversity, such as *vocd-D*, *HD-D*, and *Maas* (Fergadiotis et al., 2015; McCarthy & Jarvis, 2010; Treffers-Daller et al., 2018), so findings that are based on it are reasonably generalizable. Second, research shows that MTLD is relatively robust to short texts and to variations in text length, compared to most other measures of lexical diversity (Fergadiotis et al., 2015; Koizumi, 2012; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Vidal & Jarvis, 2020; Yan et al., 2020; Zenker & Kyle, 2021).

Another robust measure of lexical diversity that we considered is the *moving-average type–token ratio* (MATTR) (Covington & McFall, 2010; Fergadiotis et al., 2015; Vidal & Jarvis, 2020; Zenker & Kyle, 2021). We decided to focus on only one measure in the body of the article, due to length constraints. We chose MTLD over MATTR because, as noted above, there is more research on it, which facilitates the interpretation of our findings and their comparison with the findings of others, and particularly Treffers-Daller et al. (2018), who examined MTLD but not MATTR.

Nevertheless, since MATTR is calculated in a different manner than MTLD, we built models that use it as our response variable, to see if the choice of lexical-diversity measure influences our findings. MATTR is an extension of type-token ratio (TTR), which is one of the simplest measures of lexical diversity and, as introduced in the main text, is calculated by dividing the number of word types by that of word tokens. TTR is highly sensitive to text length because the rate of increase in the number of word types differs across token counts. MATTR corrects this issue by calculating multiple TTRs in a sliding-window fashion within a fixed window size and calculating their averages. Individual TTRs, therefore, are calculated with the same number of tokens throughout. Specifically, MATTR is calculated as follows:

> We choose a window length (say 500 words) and then compute the TTR for words 1–500, then for words 2– 501, then 3–502, and so on to the end of the text. The mean of all these TTRs is a measure of the lexical diversity of the entire text…
>
> (Covington & McFall, 2010, p. 96)

This means that MATTR should not theoretically suffer from the same potential issues as MTLD when it comes to assessing lexical diversity in very short texts, since it does not involve the use of lexical remainders. This is supported by Zenker and Kyle (2021), who examined the use of MATTR in texts as short as 50 tokens and found that it is fairly robust there, though they also found that MTLD was robust in those conditions.

7.4.4.4.2 Our approach

When calculating MATTR, we used the same sample as for the main models (i.e., after the removal of the small percentage of outliers), to allow for a direct comparison of the models.

We calculated MATTR using the same programmatic tools that we used to calculate MTLD. Specifically, we used the *tokenize* function in *lexical-diversity* library in Python (Kyle, 2018) to tokenize the spelling-corrected versions of the texts, and then used the library's *mattr* function to calculate the MATTR of each text, based on the approach by Covington & McFall (2010) that is outlined above.

We calculated two main sets of MATTR scores, using different window sizes: 20 and 50. The window size of 20 was used at it is the length of the shortest text in the sample, and consequently allows us to calculate a MATTR score for all the texts in the sample. However, because MATTR has not been validated using such short window size yet, we also calculated MATTR with a window length of 50, which has been previously validated (Zenker & Kyle, 2021), on all texts in our sample that had 50+ tokens.[98]

In addition, we also calculated MATTR scores with a window length of 50 using the more sophisticated TAALED-based approach to pre-processing texts (as outlined in the section of this supplementary information on our technical approach for calculating MTLD).

[98] This is not to say that MATTR is necessarily invalid when a window size of 20 is used, but rather that its validity has not yet been investigated. However, a key issue that can be expected to occur with such short window size is that there will not be enough vocabulary repetition in such short sequences for this measure to work properly. But, this does not seem to be a substantial issue in our sample, as discussed in the next sub-section, which shows that there is a very strong correlation between the two sets of MATTR scores in our sample (i.e., between those with a window size of 20 and a window size of 50). A possible reason for this is that our sample includes the writings of L2 learners with beginner to intermediate L2 proficiency, and generally, the lower a learner's L2 proficiency is, the smaller their functional vocabulary is, and the higher the rate of repetition in their productions, as we note in the section of this supplementary information discussing text length and MTLD. This also ties in to a statement made by Covington and McFall (2010, p. 97) in their paper outlining MATTR: "A short window, perhaps as short as 10 words, is appropriate if the goal is to detect repetition of immediately preceding words or phrases due to dysfluent production".

7.4.4.4.3 Analyses of MATTR values

The correlation of MTLD with MATTR scores based on a window size of 20 was very strong in both the first corpus (*r* = .79, *p* < .001, *95%CI* = [.78, .80]) and the second (*r* = .76, *p* < .001, *95%CI* = [.75, .77]).[99]

In addition, the correlation of MTLD in texts longer than 50 words with MATTR scores based on a window size of 50 was also very strong in both the first corpus (*r* = .86, *p* < .001, *95%CI* = [.85, .86]) and the second (*r* = .85, *p* < .001, *95%CI* = [.84, .86]).[100]

Similarly, the correlation of MATTR based on a window size of 20 and MATTR based on a window size of 50, in texts with 50+ tokens, was also very strong in both the first corpus (*r* = .88, *p* < .001, *95%CI* = [.87, .88]) and the second (*r* = .86, *p* < .001, *95%CI* = [.85, .87]).

Finally, the correlation between MATTR (with a window length of 50) that was calculated using simple tokenization and MATTR (again with a window length of 50) that was calculated with the more sophisticated TAALED-based pre-processing was also very strong in both the first corpus (*r* = .92, *p* < .001, *95%CI* = [.91, .92]) and the second (*r* = .88, *p* < .001, *95%CI* = [.87, .89]).

In summary, there were very strong correlations between MTLD and MATTR across the full sample (when a window size of 20 is used), as well as between MTLD and MATTR in texts with 50+ tokens (when a window size of 50 is used). In addition, there were very strong correlations between MATTR with a window size of 20 and MATTR with a window size of 50, in texts with 50+ tokens, as well as between MATTR based on a simple tokenization approach and MATTR based on a more sophisticated TAALED-based pre-processing approach (again, when a window size of 50 is used on texts with 50+ tokens).

This supports the MTLD values found in the present study—and consequently the findings based on them—by showing their strong association with the MATTR values. This

[99] There was a very small number of cases (19 in the first corpus, 0.24% of total; and 11 in the second corpus, 0.18% of total) where the number of tokens used to calculate MATTR was less than 20 (range = 16–19 in both corpora), even though the calculated wordcount in the corpus was 20 or more. This was due to minor differences in how the number of tokens is calculated by the *lexical-diversity* library compared to how it was calculated in the EFCAMDAT Cleaned Subcorpus. In such cases—where a text is shorter than the window length used to calculate MATTR, a simple TTR value is returned instead, similarly to when the text is the same length as the window. However, the influence of this is negligible here, given the very small proportion of such cases, and the relatively minor influence of this on the calculations of MATTR in these cases.

[100] We calculated this using texts which had 50+ tokens according to all three available token counts (the counts in the corpus, the counts based on the simple tokenization function, and the counts based on the TAALED-based pre-processing). In practice, this was not an issue for the absolute vast majority of texts, compared to using texts with a wordcount of 50+ in the corpus: in the first corpus, this led to the use of 6,284 texts (98.42% of those with a corpus count of 50+), and in the second corpus, this led to the use of 4,580 texts (98.05%).

also supports the use of the use of MATTR with a short window length of 20 in the present study (i.e., on the present sample, for the present analyses). Finally, this suggests that while the TAALED-based pre-processing is likely beneficial, its effect does not substantially alter our findings, as supported by the previous supplementary models (with TAALED-based MTLD), which had similar findings as the main models.

Nevertheless, these findings should be interpreted with caution, since we focused on validating the use of these measures within the context of the present study only, and as such we do not claim that these findings will necessarily generalize to other samples or analyses. Specifically, we acknowledge the potential issues associated with calculating lexical diversity (including MTLD and MATTR) using texts as short as those as the ones in our sample, and with using such a short window length in the case of MATTR. But, given the evidence outlined above, we believe that these issues do not invalidate the main findings of this study.

7.4.4.4.4 MATTR-based results

We built a set of MATTR-based models to supplement our MTLD-based ones. For this, we used the MATTR values that we calculated with a window length of 20 (using the whole sample), as well as the MATTR scores that we calculated with a window length of 50 (using a subset of the sample, including all texts from the A2–B2 CEFR level).[101,102]

The results of these models appear in Tables 36 and 37 (the assumption checks are in Figures 35, 36, 37, and 38). They largely mirror the results of the main models, and this is particularly evident when looking at the standardized coefficients, which can be directly compared with the standardized coefficients in the main (MTLD-based) models.

Specifically, across all models, there is a no effect of lexical distance, and no interaction between it and L2 proficiency.[103] In addition, there is a similarly significant association

[101] We used the same subset of the sample as for MTLD-based supplementary models that are described in the section on text length and MTLD in the supplementary information. As shown in the descriptive statistics in that section, the vast majority of the texts in this subset had a wordcount equal to or greater than 50. To avoid a selection bias—which is also the reason why we focused here on texts from A2–B2, rather than just texts with 50+ words—we included all the texts from those proficiency levels in our analyse. 96.95% in the first corpus and 91.80% in the second corpus contained 50+ tokens (based on the number of tokens used to calculate the MATTR scores).

[102] Since the TAALED-based MATTR values were so strongly correlated with those based on the simpler pre-processing approach, and since using the TAALED based values did not make a difference for the MTLD-based models, we did not build TAALED-based models here, as they would be redundant.

[103] In terms of significance, the interaction term in the second corpus is borderline significant based on the criterion of p = .05, but this appears to spurious, given (1) that its p-value is only borderline significant (p = .043), (2) the effect size is low enough to be negligible, (3) the interaction is non-significant in the other corpus, (4) the effect

between L2 proficiency and lexical diversity, which appears when looking at the full sample, but weakens or disappears when looking only at texts at the A2–B2 range. Finally, when it comes to the random effects, there is again a relatively large effect of *task* and a negligible effect of *L1*. This similarity in results across the MTLD- and MATTR-based models is expected, given the very strong correlations between MTLD and MATTR presented in the previous sub-section.

Overall, the supporting models that use MATTR as a measure of lexical diversity show very similar results as the MTLD-based models. This provides strong support for the main findings, especially given that MTLD and MATTR are the two main lexical-diversity measures that are recommended for use when it comes to the kind of learner data that was used here, and given the strong correlation between these two measures and between these measures and other lexical-diversity measures, such as *vocd-D*, *HD-D*, and *Maas* (Fergadiotis et al., 2015; Koizumi, 2012; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Treffers-Daller et al., 2018; Yan et al., 2020; Zenker & Kyle, 2021).

of distance is insignificant and insubstantial in both corpora, and (5) we are making multiple comparisons, so it is expected that there will be some borderline cases such as this simply due to chance.

Table 36. Results of the models that were calculated with MATTR as a response variable instead of MTLD, using a window length of 20. MATTR was scaled by a factor of 100 (so a scale of 0–100 instead of 0–1), to facilitate the interpretation of the unstandardized coefficients as well as comparison with MTLD.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	87.00	0.34	86.33 – 87.68	<0.001	0.01	-0.10 – 0.12	88.24	0.31	87.63 – 88.84	<0.001	-0.00	-0.12 – 0.11
Lexical_distance	-1.21	1.95	-5.04 – 2.62	0.537	-0.01	-0.06 – 0.03	-0.43	2.01	-4.37 – 3.52	0.833	-0.01	-0.06 – 0.05
L2_proficiency	0.89	0.09	0.71 – 1.07	<0.001	0.49	0.39 – 0.59	0.60	0.08	0.44 – 0.75	<0.001	0.40	0.30 – 0.51
Lexical_distance * L2_proficiency	0.13	0.21	-0.28 – 0.54	0.536	0.01	-0.01 – 0.02	0.44	0.22	0.01 – 0.87	0.043	0.02	0.00 – 0.04
Random Effects												
σ^2	17.40						13.53					
τ_{00}	2.95_{Learner}						3.86_{Learner}					
	9.28_{Task}						5.17_{Task}					
	0.16_{L1}						0.17_{L1}					
ICC	0.42						0.40					
N	9_{L1}						9_{L1}					
	5385_{Learner}						4357_{Learner}					
	95_{Task}						71_{Task}					
Observations	8081						6129					
Mar. R^2 / Cond. R^2	0.239 / 0.555						0.159 / 0.499					

Table 37. Results of the models that were calculated with MATTR as a response variable instead of MTLD, using a window length of 50, and texts at the A2–B2 CEFR range. MATTR was scaled by a factor of 100 (so a scale of 0–100 instead of 0–1), to facilitate the interpretation of the unstandardized coefficients as well as comparison with MTLD.

	First corpus						Second corpus					
Predictors	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*	*B*	*SE*	*95% CI*	*p*	*std. B*	*std. 95% CI*
(Intercept)	74.57	0.54	73.51 – 75.64	<0.001	0.00	-0.14 – 0.14	77.89	0.49	76.92 – 78.85	<0.001	0.00	-0.15 – 0.16
Lexical_distance	0.88	2.98	-4.95 – 6.72	0.767	0.00	-0.06 – 0.07	2.53	3.09	-3.53 – 8.59	0.413	0.03	-0.05 – 0.11
L2_proficiency	1.01	0.17	0.68 – 1.34	<0.001	0.38	0.26 – 0.51	0.08	0.15	-0.21 – 0.37	0.599	0.04	-0.10 – 0.17
Lexical_distance * L2_proficiency	-0.46	0.36	-1.16 – 0.25	0.203	-0.01	-0.03 – 0.01	-0.24	0.39	-0.99 – 0.52	0.542	-0.01	-0.03 – 0.02
Random Effects												
σ^2	23.66						18.67					
τ_{00}	3.05_{Learner}						4.66_{Learner}					
	13.31_{Task}						7.61_{Task}					
	0.39_{L1}						0.41_{L1}					
ICC	0.41						0.40					
N	9_{L1}						9_{L1}					
	3882_{Learner}						3182_{Learner}					
	71_{Task}						53_{Task}					
Observations	6160						4621					
Mar. R^2 / Cond. R^2	0.145 / 0.499						0.002 / 0.406					

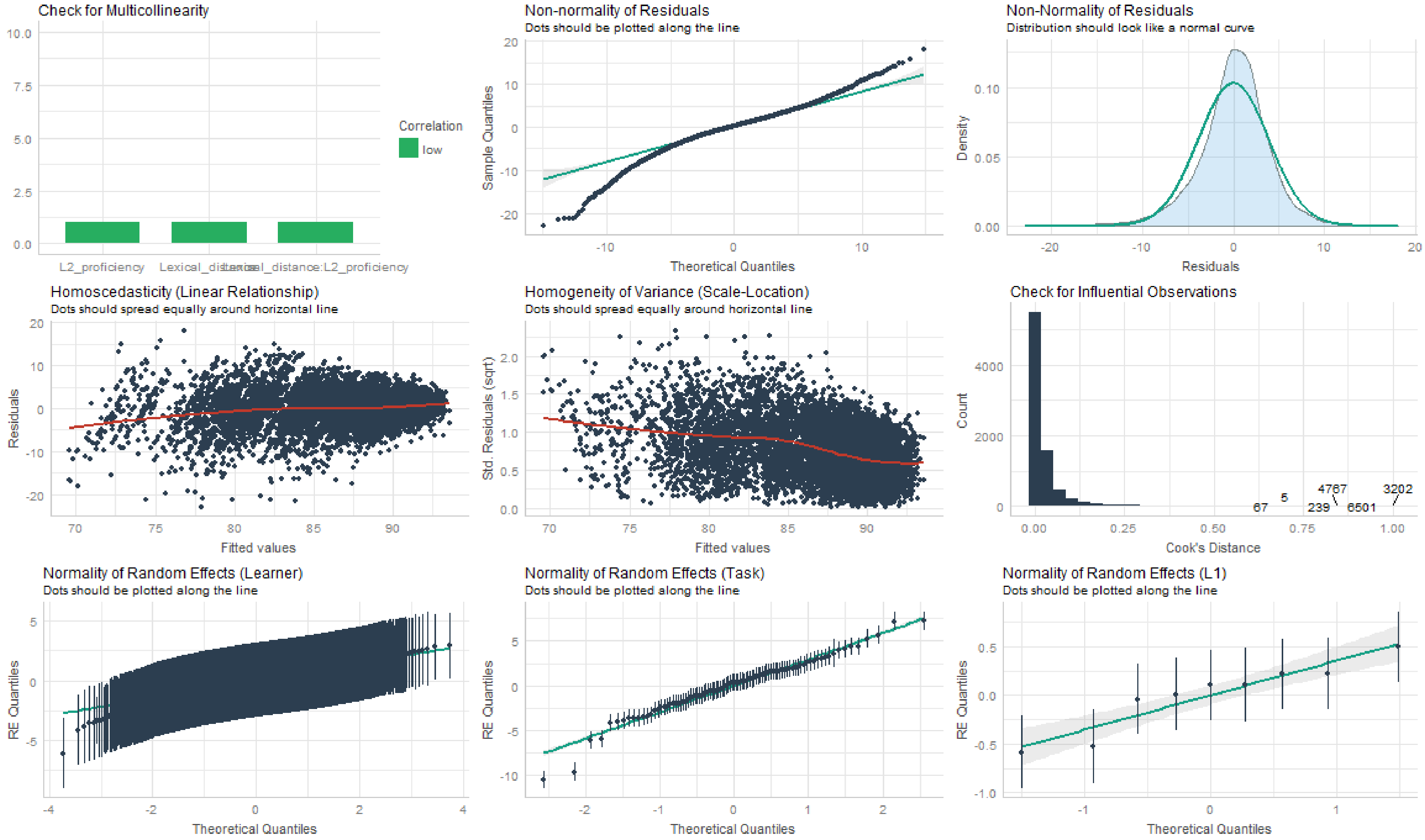


Figure 35. Assumption checks for the model MATTR as the response variable (using a window length of 20), in the first corpus.

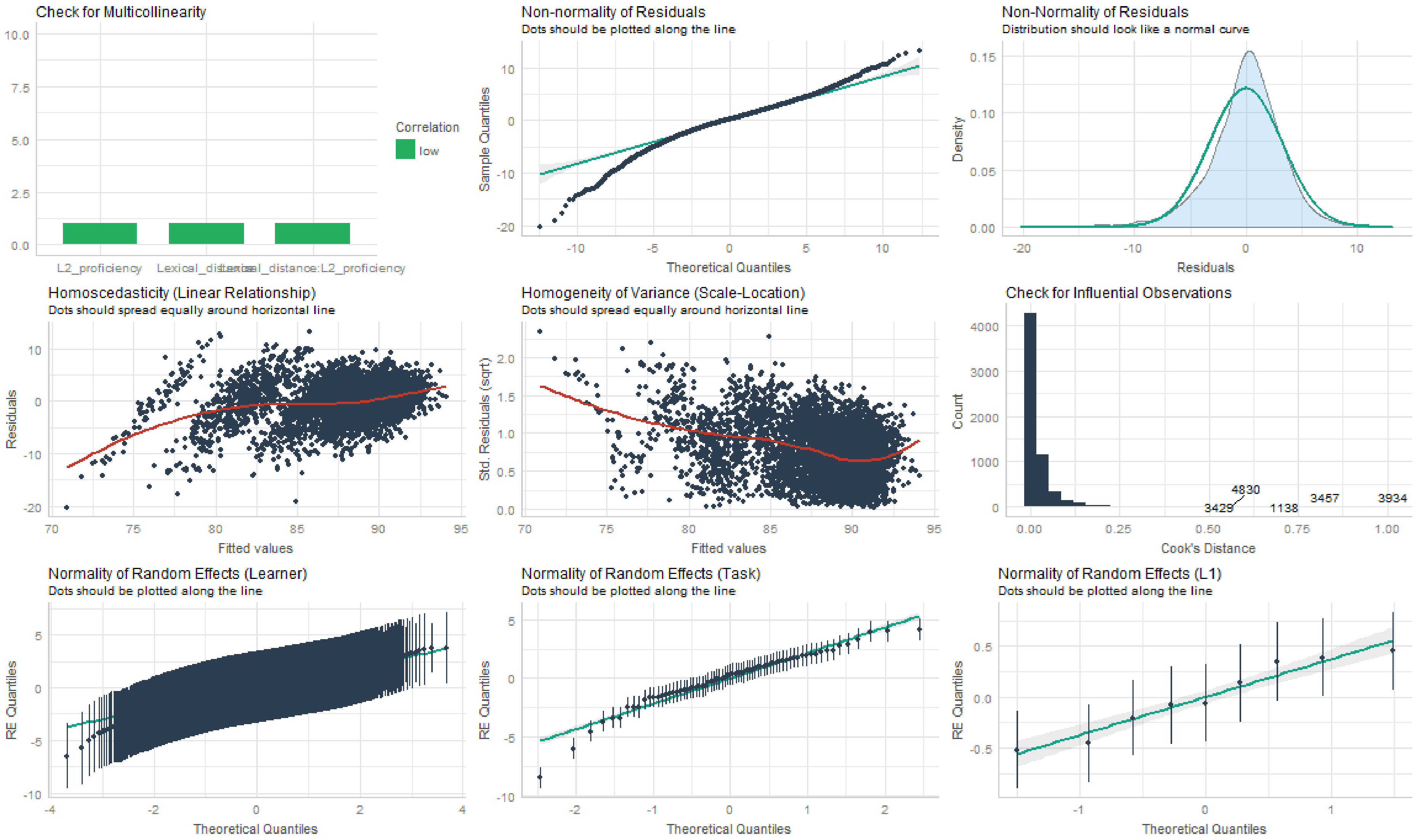


Figure 36. Assumption checks for the model MATTR as the response variable (using a window length of 20), in the second corpus.

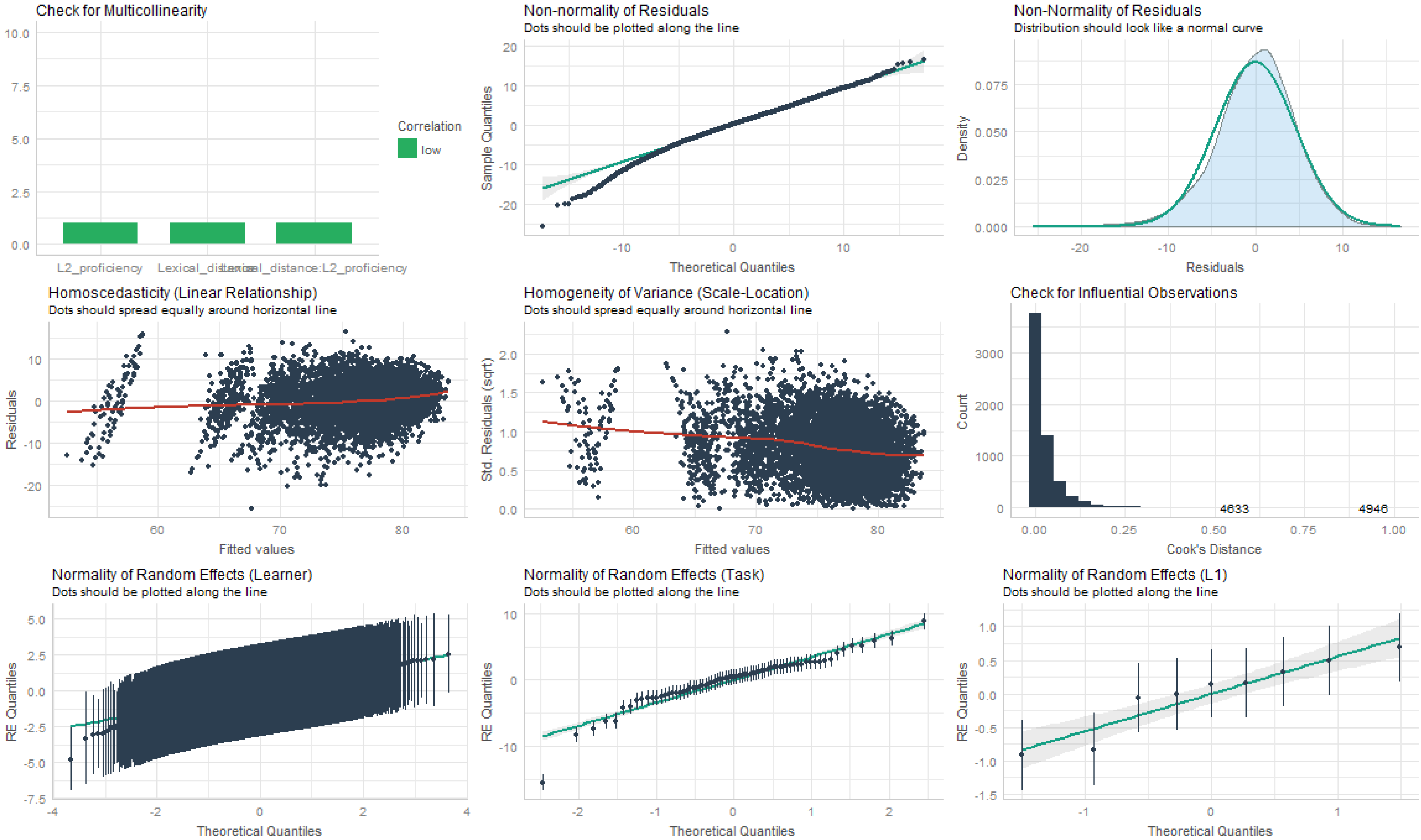


Figure 37. Assumption checks for the model MATTR as the response variable (using a window length of 50 and texts at the A2–B2 CEFR range), in the first corpus.

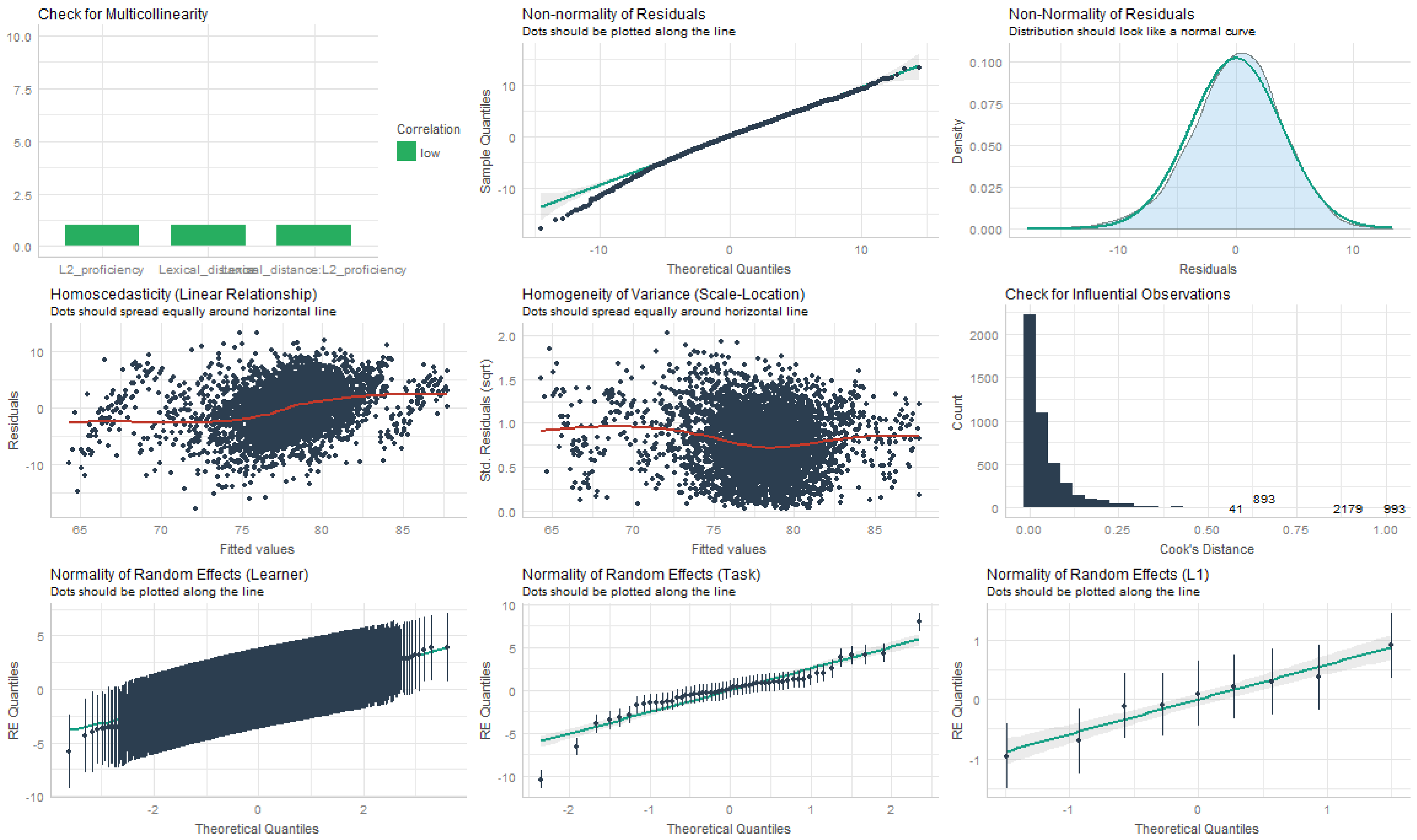


Figure 38. Assumption checks for the model MATTR as the response variable (using a window length of 50 and texts at the A2–B2 CEFR range), in the second corpus.

7.4.5 Supporting statistics

7.4.5.1 Raw correlations of lexical diversity and proficiency

The *Pearson's r* correlation between lexical diversity (MTLD) and L2 proficiency (based on EFCAMDAT level) was $r = .51$, $p < .001$, *95%CI* = [.49, .52] in the first corpus and $r = .43$, $p < .001$, *95%CI* = [.41, .45] in the second corpus. *Spearman's rho* correlations are very similar (.52 in the first corpus and .45 in the second, both $p < .001$). Likewise, so are the correlations when using task number as the measure of proficiency ($r = .51$, $p < .001$, *95%CI* = [.49, .52] in the first corpus and $r = .44$, $p < .001$, *95%CI* = [.41, .46] in the second), and when using CEFR as the measure of proficiency ($r = .49$, $p < .001$, *95%CI* = [.48, .51] in the first corpus and $r = .44$, $p < .001$, *95%CI* = [.42, .46] in the second).

7.4.5.2 Lexical diversity per CEFR level, by L1

Table 38 contains details about lexical diversity per CEFR level and by L1 in each corpus.

Table 38. The mean and the standard deviation of lexical diversity (MTLD), per CEFR level and by L1 in each corpus.

L1	CEFR level	Mean (SD) [first corpus]	Mean (SD) [second corpus]
Arabic	A1	43.36 (20.03)	45.16 (18.75)
Arabic	A2	56.01 (23.09)	63.54 (26.23)
Arabic	B1	71.80 (22.52)	72.97 (24.03)
Arabic	B2	73.44 (22.99)	76.17 (19.73)
French	A1	45.59 (21.21)	46.96 (16.85)
French	A2	58.01 (20.61)	69.21 (26.02)
French	B1	74.18 (22.46)	76.45 (23.78)
French	B2	80.50 (23.20)	78.07 (21.59)
German	A1	46.35 (20.28)	46.63 (15.78)
German	A2	58.15 (22.44)	67.52 (23.38)
German	B1	73.58 (23.52)	75.59 (21.56)
German	B2	80.01 (24.40)	76.92 (19.74)
Italian	A1	47.10 (20.42)	45.85 (15.55)
Italian	A2	60.52 (24.58)	66.94 (23.60)

L1	CEFR level	Mean (SD) [first corpus]	Mean (SD) [second corpus]
Italian	B1	76.66 (23.55)	75.97 (22.96)
Italian	B2	80.36 (23.81)	74.60 (19.95)
Japanese	A1	45.58 (20.71)	48.26 (17.72)
Japanese	A2	57.82 (21.80)	68.76 (27.14)
Japanese	B1	74.86 (23.16)	74.40 (25.10)
Japanese	B2	78.68 (26.03)	77.97 (20.95)
Mandarin	A1	45.32 (20.74)	48.91 (19.08)
Mandarin	A2	61.77 (23.83)	73.85 (24.08)
Mandarin	B1	75.18 (21.88)	78.59 (22.94)
Mandarin	B2	79.45 (22.88)	79.18 (22.12)
Portuguese	A1	46.28 (20.06)	44.84 (16.07)
Portuguese	A2	58.10 (22.57)	68.93 (22.70)
Portuguese	B1	75.16 (21.31)	73.33 (24.87)
Portuguese	B2	80.02 (24.92)	74.62 (22.09)
Russian	A1	47.10 (20.47)	48.18 (17.68)
Russian	A2	58.94 (22.43)	67.88 (26.62)
Russian	B1	79.49 (22.97)	80.50 (22.45)
Russian	B2	86.41 (26.31)	81.12 (21.38)
Spanish	A1	43.04 (20.38)	44.44 (16.23)
Spanish	A2	57.13 (22.35)	63.37 (22.46)
Spanish	B1	69.79 (21.15)	70.68 (22.84)
Spanish	B2	72.88 (23.28)	71.22 (18.39)

7.4.5.3 Number of texts per CEFR level, by L1

Table 39 contains the number of texts per CEFR level by L1 in each corpus, after removing outliers.

Table 39. Number of texts in the final sample after the removal of outliers, per L1 and CEFR proficiency level.

L1	First corpus					Second corpus				
	Total	A1	A2	B1	B2	Total	A1	A2	B1	B2
Arabic	863	206	233	226	198	688	171	172	167	178
French	911	222	230	229	230	681	162	176	167	176
German	903	211	231	229	232	691	170	174	170	177
Italian	905	215	232	226	232	688	174	172	166	176
Japanese	901	217	231	227	226	677	168	172	165	172
Mandarin	887	210	230	222	225	660	158	168	164	170
Portuguese	903	208	234	229	232	683	164	174	168	177
Russian	908	216	233	227	232	669	168	170	164	167
Spanish	900	216	234	227	223	692	173	175	169	175
Total	*8,081*	*1,921*	*2,088*	*2,042*	*2,030*	*6,129*	*1,508*	*1,553*	*1,500*	*1,568*

7.4.6 Information about the analyses in R

All analyses were conducted in R (R Core Team, 2021).[104] All tests of statistical significance throughout the study were two-tailed. The mixed-effects models were built using the *lmer* function in the *lme4* package in R (Bates et al., 2015), using the default settings (*REML* and a *nloptwrap* optimizer). Assumption checks were generated using the *performance* package in the *easystats* ecosystem (Lüdecke et al., 2021). The scatterplots and linear models with lexical distance as the predictor and lexical diversity as the response variable (per CEFR level) were generated in *ggplot2* using the *lm* method under *geom_smooth* (Wickham, Averick, et al., 2019).

To list the specific packages that were loaded throughout the analyses, we used the *sessionInfo* function from the *report* library (Makowski & Lüdecke, 2019). Note that this generates an automated output based on the citation information associated with the metadata of each package, which may be incomplete or formatted differently than APA style. We present this bibliography here as-is, to preserve the original output, and we therefore also separate it from the main the main bibliography for this document.

---Start of report(sessionInfo()) output below---

Analyses were conducted using the R Statistical language (version 4.0.4; R Core Team, 2021) on Windows 10 x64 (build 19042), using the packages ggpubr (version 0.4.0; Alboukadel Kassambara, 2020), cowplot (version 1.1.1; Claus Wilke, 2020), Matrix (version 1.3.2; Douglas Bates and Martin Maechler, 2021), lme4 (version 1.1.26; Douglas Bates et al., 2015), Hmisc (version 4.5.0; Frank E Harrell Jr, with contributions from Charles Dupont and many others., 2021), ggplot2 (version 3.3.3; Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.), stringr (version 1.4.0; Hadley Wickham, 2019), forcats (version 0.5.1; Hadley Wickham, 2021), tidyr (version 1.1.3; Hadley Wickham, 2021), readr (version 1.4.0; Hadley Wickham and Jim Hester, 2020), dplyr (version 1.0.5; Hadley Wickham et al., 2021), tibble (version 3.1.0; Kirill Müller and Hadley Wickham, 2021), purrr (version 0.3.4; Lionel Henry and Hadley Wickham, 2020), ggeffects (version 1.0.2; Lüdecke D, 2018),

[104] There are two exceptions to this. First, MATTR was calculated using Python, as noted in the section on MATTR in this document. Second, the lexical-distance data was also generated in Python. Specifically, the following Python libraries were used for some basic data wrangling and calculations: *SciPy* (Virtanen et al., 2019), *pandas* (McKinney, 2010), and *numpy* (Oliphant, 2006; Walt et al., 2011). The ASJP’s specialized phonetic script (outlined in Brown et al., 2008) was converted to IPA using the dedicated *asjp* library (Sofroniev, 2018). The distances were calculated using the *PanPhon* library (Mortensen et al., 2016).

sjPlot (version 2.8.7; Lüdecke D, 2021), performance (version 0.7.0; Lüdecke et al., 2020), report (version 0.2.0; Makowski et al., 2020), data.table (version 1.14.0; Matt Dowle and Arun Srinivasan, 2021), openxlsx (version 4.2.3; Philipp Schauberger and Alexander Walker, 2020), janitor (version 2.1.0; Sam Firke, 2021), lattice (version 0.20.41; Sarkar, Deepayan, 2008), survival (version 3.2.7; Therneau T, 2020), tidyverse (version 1.3.0; Wickham et al., 2019) and Formula (version 1.2.4; Zeileis A, Croissant Y, 2010).

References

- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

- Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

- Douglas Bates and Martin Maechler (2021). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.3-2. https://CRAN.R-project.org/package=Matrix

- Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

- Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2021). Hmisc: Harrell Miscellaneous. R package version 4.5-0. https://CRAN.R-project.org/package=Hmisc

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

- Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

- Hadley Wickham (2021). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. https://CRAN.R-project.org/package=forcats

- Hadley Wickham (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

- Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr

- Kirill Müller and Hadley Wickham (2021). tibble: Simple Data Frames. R package version 3.1.0. https://CRAN.R-project.org/package=tibble

- Lionel Henry and Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr

- Lüdecke D (2018). “ggeffects: Tidy Data Frames of MarginalEffects from Regression Models.” _Journal of Open SourceSoftware_, *3*(26), 772. doi: 10.21105/joss.00772 (URL:https://doi.org/10.21105/joss.00772).

- Lüdecke D (2021). _sjPlot: Data Visualization for Statisticsin Social Science_. R package version 2.8.7, <URL:https://CRAN.R-project.org/package=sjPlot>.

- Lüdecke, Makowski, Waggoner & Patil (2020). Assessment of Regression Models Performance. CRAN. Available from https://easystats.github.io/performance/

- Makowski, D., Lüdecke, D., & Ben-Shachar, M.S. (2020). Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption. CRAN. Available from https://github.com/easystats/report. doi: .

- Matt Dowle and Arun Srinivasan (2021). data.table: Extension of `data.frame`. R package version 1.14.0. https://CRAN.R-project.org/package=data.table

- Philipp Schauberger and Alexander Walker (2020). openxlsx: Read, Write and Edit xlsx Files. R package version 4.2.3. https://CRAN.R-project.org/package=openxlsx

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

- Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

- Therneau T (2020). _A Package for Survival Analysis in R_. Rpackage version 3.2-7, <URL:https://CRAN.R-project.org/package=survival>.

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

- Zeileis A, Croissant Y (2010). “Extended Model Formulas in R:Multiple Parts and Multiple Responses.” _Journal ofStatistical Software_, *34*(1), 1-13. doi:10.18637/jss.v034.i01 (URL:https://doi.org/10.18637/jss.v034.i01).

---*End of report(sessionInfo()) output above*---

7.5 Appendix E: Supplementary information for study 2 (on word choice)

7.5.1 Lexical distance

7.5.1.1 The term "lexical distance"

There is no universal distinction between the terms *lexical distance* and *lexical similarity*, which are often used interchangeably.[105] In the present study, we use *lexical distance* to refer to distance between individual L1-L2 words, which are translations of one another. This distance is based on objective phonological distance (specifically, normalized Levenshtein distance—LDN), which serves as a proxy for the subjective similarity between the words that is expected to be perceived by speakers, as supported by the studies outlined in the next sub-section (§7.5.1.2).

The reason we use the term *lexical distance* in particular is to distinguish it from other types of language distances, such as morphological distance, in line with prior studies (Bakker et al., 2009; Brown et al., 2008; Gooskens, 2006; Holman et al., 2008b; Schepens et al., 2016; Schepens, van der Slik, et al., 2013b). Here, it is worth noting that lexical distance can serve as a proxy of overall *language distance*, which is sometimes also referred to as *linguistic distance* or *typological distance* (e.g., Ecke, 2015; Llach, 2010),[106] but we do not use it in this sense here, since in our study we only consider the distance of individual L1-L2 word pairs directly, rather than the distance between languages as a whole. Nevertheless, note that we are using a specific type of lexical distance—phonological LDN—as a proxy for overall lexical distance, which can include other factors, such as orthography.[107] Finally, note that other less-common

[105] Though *increased* distance denotes *decreased* similarity and vice versa, so lexical distance is technically more closely associated with lexical *dissimilarity*.

[106] Though one issue with the term "typological distance" is that it is not always used to refer to overall language distance. Rather, it is sometimes used to refer to distance that is based on grammatical features, such as those that are available in the *World Atlas of Language Structures* (WALS), in order to draw a distinction between it and other types of distance, such as lexical distance that is based on Levenshtein distance in Swadesh lists (Bakker et al., 2009).

[107] Though note that phonological and orthographic similarity tend to be highly correlated. For example, in a recent study on the English and French vocabulary of Dutch speaking children, De Wilde et al.(2021), who also used normalized Levenshtein distance, included only phonological similarity in their analyses, and omitted orthographic similarity, since the two variables were highly correlated and could therefore lead to issues with collinearity. This is also something that these researchers did in another associated study (De Wilde et al., 2020), and is an issue that was raised by other researchers, such as Carrasco-Ortiz et al. (2021), who found a correlation of $r = .782$ between orthographic and phonological distance in their dataset of English and Spanish words. Here too, there were similarly strong correlations between phonological and orthographic distance in the parallel dictionaries, where all the L1s share English's Latin script, both in the case of LDN ($r = .68$, *95% CI* = [.67, .70], $p < .001$), and in the case of LD ($r = .73$, *95% CI* = [.71, .74], $p < .001$). We do not include orthographic distance in our analyses both because of its substantial overlap with phonological similarity in the case of L1s that share English's script, and because we wanted to use consistent analyses for all the L1s in the sample, but orthographic distance is largely meaningless across languages that use different scripts (which includes several of the L1s in the Swadesh-based sample).

terms are sometimes used for lexical distances that are similar to the one that we use here, such as *phonological overlap* (Carrasco-Ortiz et al., 2021) or *cognate linguistic distance* (van der Slik, 2010).[108]

7.5.1.2 Validation of Levenshtein distance

Here, we outline the extensive use and validation that Levenshtein distance (LD) and its normalized form (LDN) have received in prior research.[109]

First, we open with a notable study by Schepens et al. (2012), which is often cited by other researchers in this context (e.g., Blom et al., 2020; Cenoz et al., 2021; Cop et al., 2017; De Wilde et al., 2020, 2021; Otwinowska & Szewczyk, 2019; Silveira & Leussen, 2015; Wieling et al., 2014). Specifically, in their study, Schepens et al. (2012) conclude the following:

> It is possible to automatically identify large distributions of cognates with respect to form-similarity in various European languages by means of a formalized form-similarity metric such as normalized Levenshtein distance. Applying this metric to a professional translation database, similarity norms were obtained that are comparable to experimentally acquired orthographic similarity ratings (Dijkstra et al., 2010; Tokowicz et al., 2002), and lead to high correlations (around .90) and a large proportion of correctly classified stimuli (over 90%). The obtained distributions were also compared to an account of cross-language similarity based on Gray and Atkinson (r = .72). A common pattern in the degree of orthographic similarity of these distributions was observed within languages of the same family. In our analysis, English showed characteristics of multiple language families (Germanic, Romance). Cognate distributions were computed here using semi-complete lexicons, whereas Gray and Atkinson used only a small set of high frequency words.
>
> In all, our study demonstrated the feasibility and advantages of applying techniques from artificial intelligence to psycholinguistic and linguistic research involving multiple languages. First, the application of the normalized Levenshtein distance function resulted in an automatized selection of more and better stimulus materials for cognate studies on bilingual word processing. Second, the Levenshtein distance

[108] Though the term "cognate lexical distance" is not appropriate to use here, since it refers to the overall distance between languages as calculated based on the proportion of cognates, rather than to distances that are calculated between individual word pairs.

[109] Note that LDN is sometimes also referred to in the literature using similar terms, and especially *NLD* and *nLD*.

> function yielded accurate and detailed cross-language similarity distributions for multiple languages, thus allowing a comparison to language family trees. As such, the present study has shown that the Levenshtein distance function can compete with existing similarity measures (such as those proposed by Coltheart, Davelaar, Jonasson & Besner, 1977, and Van Orden, 1987) and can be considered as a new formal and computational model of orthographic similarity, useful for future empirical studies in monolingual and bilingual domains as diverse as those dealing with neighborhood effects, spelling systems, and dyslexia.
>
> (p. 165)

In addition, further support for LD(N) as a measure of lexical distance comes from many other studies.

First, there is substantial support for this measure based on its extensive use in studies pertaining to language classification. For example, in a study that examined lexical distance between 35 Indo-European L1s and Dutch, Schepens et al. (2013b) found a very high correlation (*r* = .90) between this measure as determined based on the ASJP's Swadesh lists, and distances that are based on shared cognates as determined by Gray and Atkinson (2003) on historical-comparative grounds. Furthermore, Schepens et al. (2013a) found that this measure correlates strongly with crosslinguistic morphological similarity (*r* = -.65), as determined based on morphological features in the World Atlas of Language Structures. In addition, based on comparisons with other data sources, such as established dialect boundaries, using LD between phonetic strings has been shown to be effective for assessing dialects, for example when it comes to Gaelic (Kessler, 1995) and Dutch (Gooskens & Heeringa, 2004; Nerbonne & Heeringa, 2001). Finally, other studies have found that this measure leads to accurate language classification as determined based on measures such as expert classification, when it comes to many other languages and dialects (Schepens et al., 2012; Serva & Petroni, 2008; Wichmann et al., 2010).

There is also substantial support for LD(N) based on the high correlation between it and various psycholinguistic measures (Heeringa & Prokić, 2018).[110] For example, Beijering et al. (2008) found a strong correlation between LDN-based distances and intelligibility scores

[110] This is important, since LD/LDN are *objective* measures of language distance, which often serve—including in the present research—as proxies for the *subjective* language distance that learners perceive (i.e., the *psychotypology*), which is the main driver behind the crosslinguistic influence that they experience (Jarvis & Pavlenko, 2008; Kellerman, 1983; Ringbom, 2007; Xia, 2017).

(*r* = -.86) and perceived linguistic distances (*r* = .52), in their study of Standard Danish and 17 other Scandinavian language varieties.[111] Similarly, Gooskens (2006) found a correlation of *r* = -.82 between phonetic LD and intelligibility scores among students from schools in Denmark, Norway, Sweden, and Finland. Furthermore, Gooskens and Heeringa (2004), who examined 15 Norwegian dialects as judged by Norwegian listeners, found a strong correlation between LD and perceptual distance (*r* = .62 in an experiment where monotonized recordings were used, and *r* = .67 in an experiment where nonmanipulated recordings were used), leading the researchers to conclude that:

> This shows that dialect distances calculated with Levenshtein distance approximate perceptual distances rather well. We see this as a confirmation of the usefulness of the Levenshtein method, as has been shown before for Dutch dialects. Now we know that the method is also applicable in a language area with a less simple geographic situation than the Dutch one.
>
> (p. 205)

Furthermore, this measure has also been extensively used and validated in the context of second language acquisition (SLA) research, which involved similar analyses as the present study. This includes the following studies:

- Otwinowska et al. (2020) used LDN to quantify L1-L2 orthographic similarity between words, in their study on the influence of cross-linguistic lexical similarity on the learning of cognates and non-cognates among Polish learners of English. Specifically, they used this measure to show that the cognates and false cognates that they examined were comparable in terms of their L1-L2 orthographic similarity, and this measure has been used in similar ways in associated studies (e.g., Marecka et al., 2021; Otwinowska & Szewczyk, 2019).[112]
- Many studies used this measure to assess cognancy. This includes using LD to determine cognancy based on phonological (Sadat et al., 2016) or orthographic transcriptions (Bultena et al., 2020; Y. Zhu & Mok, 2020), using LD to compare cognates and non-cognates based on both phonological and orthographic transcriptions (Carrasco-Ortiz et al., 2021), using LDN to determine cognancy based on orthographic

[111] They also found similar correlations when it comes to non-normalized LD (*r* = -.79 for intelligibility and *r* = .62 for perceived distance).

[112] Otwinowska et al. (2020) also used LDN to quantify orthographic dissimilarity "between a correct L2 translation and a participant's response that was required to treat the response as correct" (p. 712), and other researchers used this measure for similar comparative purposes (Hanulíková et al., 2012; Marecka et al., 2021).

transcriptions (Casaponsa et al., 2015), and using LDN to determine cognancy based on both phonological and orthographic transcriptions (De Wilde et al., 2020).

- In addition, LD/LDN were also used in other studies to assess crosslinguistic similarity of words and its influence on L2 acquisition (De Wilde et al., 2021; van de Ven et al., 2019), to quantify crosslinguistic orthographic overlap of non-identical cognates (Vanlangendonck et al., 2020), and to serve various similar purposes (Cenoz et al., 2021), as have other closely related measures of lexical distance (Dijkstra et al., 2010; Schepens, van der Slik, et al., 2013a).

Note that many of the aforementioned SLA studies also found that this measure of lexical distance is an accurate predictor of various L2 outcomes, including L2 meaning recognition (De Wilde et al., 2021), word processing speed and accuracy (Casaponsa et al., 2015), word recognition (Carrasco-Ortiz et al., 2021), receptive word knowledge (De Wilde et al., 2020), word retrieval (Sadat et al., 2016), translation accuracy (van de Ven et al., 2019), increased errors in the case of gender-incongruent cognates (Bultena et al., 2020), and overall L2 proficiency (Schepens, van der Slik, et al., 2013a).

Finally, in the case of the present study, the classification of L1s based on their lexical distance from English aligns with what we expect based on general language classification. Specifically, based on the distances per L1, which are shown in Table 40, the Germanic and Romance L1s are the lexically closest to English, and all the Indo-European L1s are closer to English than all the non-Indo-Eurpoean L1s (Eberhard et al., 2021).

Table 40. The lexical distances between each L1 and English. This is based on the Swadesh lists, since they contain data for all the L1s in the present sample, and specifically on the data *before* the removal of multi-word entries, unlike the similar table in the body of the paper. The reason for this is that the inclusion of only single-word entries is appropriate for the analyses of individual word pairs, and therefore does not interfere with our main analyses, but could bias comparisons at the language level, where it is important to include all the available word pairs.

			Lexical distance	
L1	Language family [a]	Indo-European [a]	mean	SD
German	Germanic	Y	.656	.27
Italian	Romance	Y	.820	.20
Spanish	Romance	Y	.840	.20
French	Romance	Y	.851	.19
Russian	Slavic	Y	.867	.19
Portuguese	Romance	Y	.878	.18
Japanese	Japonic	N	.892	.15
Arabic	Semitic	N	.912	.12
Mandarin	Sino-Tibetan	N	.920	.12

Note. These values are calculating using English-based tables, where distances are calculated from each English word in the dataset to its closest L1 synonym. It is also possible to calculate these distances using L1-based tables, where distances are calculated from each L1 word to its closest English synonym. However, the distances are quite similar regardless of which option is used (*Spearman's ρ* = 0.97, *p* < .001); the key differences are that when L1-based tables are used, the Spanish-English distance increases to make it more distant than French, and the Russian-Portuguese distance increases to make it more distant than Portuguese.

[a] Language classifications are based on (Eberhard et al., 2021).

The fact that the Indo-European L1s were found to be lexically closer to English also aligns our expectations based on the measure of linguistic distance proposed by Chiswick and Miller (2005). Specifically, this measure is based on the difficulty that English speakers have acquiring other languages, and has been shown by Chiswick and Miller to predict the difficulty that speakers of those languages will have when acquiring English as an L2. Similarly to our measure of distance, their measure also suggests that all the Indo-European L1s that are included here are closer to English than the non-Indo-Eurpoean L1s.[113] Furthermore, in this

[113] Their measure ranks languages on a scale of 1–3, where 1 marks the hardest languages to learn (i.e., the most distant) and 3 marks the easiest languages to learn (i.e., the least distant). Out of the L1s included in the present sample, French, Italian, and Portuguese have a ranking of 2.5, German, Spanish, and Russian, have a ranking of 2.25, Arabic and Mandarin have a ranking of 1.5, and Japanese has a ranking of 1. This roughly corresponds to the ranking found here, whereby all the Indo-European L1s are closer to English than the non-Indo-European L1s. The imperfect correlation between their measure and ours is expected, since, as they note, their measure includes various aspects of the language beyond vocabulary, such as syntax.

regard, the use of our measure of lexical distance is further supported by Schepens et al. (2013a), who calculated lexical distance in a similar manner as us between 49 L1s and Dutch, and found that increased distance is strongly correlated (*r* = -.80) with broad L2 proficiency in Dutch.[114] This suggests that distances that are based on this measure strongly predict L2 learnability, in a similar manner as proposed by Chiswick and Miller.

In summary, there is extensive support for our use of LDN as a measure of lexical distance here, including in terms of construct and convergent validity. This includes:

- Many studies that validated it by comparing it to other measures of language classification, such as expert cognancy judgments (Brown et al., 2008; Gooskens & Heeringa, 2004; Holman et al., 2008b; Kessler, 1995; Nerbonne & Heeringa, 2001; Schepens et al., 2012; Schepens, van der Slik, et al., 2013b, 2013a; Serva & Petroni, 2008; Wichmann et al., 2010).
- Many studies that validated it by comparing it to psycholinguistic measures of language perception, such as perceived distance (Beijering et al., 2008; Gooskens, 2006; Gooskens & Heeringa, 2004; Heeringa & Prokić, 2018).
- Many SLA studies that used it for similar purposes, to assess crosslinguistic similarity (particularly cognancy), and found that it predicts many types of L2 outcomes (Bultena et al., 2020; Carrasco-Ortiz et al., 2021; Casaponsa et al., 2015; Cenoz et al., 2021; De Wilde et al., 2020, 2021; Hanulíková et al., 2012; Marecka et al., 2021; Otwinowska et al., 2020; Otwinowska & Szewczyk, 2019; Sadat et al., 2016; Schepens, van der Slik, et al., 2013a; van de Ven et al., 2019; Vanlangendonck et al., 2020; Y. Zhu & Mok, 2020).
- The alignment of the overall crosslinguistic lexical distances in our samples with what is expected based on general language classification.

That said, this measure, like all linguistic measures, is imperfect, and we recommend that future work replicate our analyses using other distance measures,[115] as we do ourselves using feature edit distance. Furthermore, it is important to remember that the validation of this measure is itself imperfect, in the sense that the studies that validated it likely had their own limitations and shortcomings, and their methodologies and goals do not always align with our own.

[114] Schepens et al. base this on distances as calculated using Swadesh lists in the ASJP, similarly to us, though they use LDND rather than LDN; this is a closely associated variant of Levenshtein distance, which is discussed in detail in the next sub-section.

[115] For more information on the issues with this measure, see the “Limitations of LDN” sub-section in the paper’s methodology. Also, additional criticism of this measure—primarily in the context of language classification—can be found in Greenhill (2011).

Nevertheless, given all the support for this measure outlined above, we believe that its use here is reasonable, and that the outcomes based on it are reasonably reliable and generalizable.

7.5.1.3 LDN vs. LDND

As noted in the body of the paper, LDN is the normalized version of LD, which accounts for word length by diving the LD between a pair or words by the length of the longer word, to control for variations in word length.

LDN can be further normalized into *LDND*, by dividing it by the mean LDN of all N(N-1)/2 pairings of words with different meanings, to control for shared phonotactic preferences or overlap in phoneme inventories (Bakker et al., 2009, p. 171). However, while the first normalization of LD is usually seen as crucial, the second normalization is controversial and rare (Petroni & Serva, 2010; Wichmann et al., 2010), and none of the SLA or psycholinguistic studies outlined in the previous sub-section (§7.5.1.2) used it. Furthermore, the use of LDND can lead to two notable issues. First, it is not sample-independent unlike LDN, so the LDND between two words varies based on which others words from the same languages are included in the analysis, which is not the case for LDN. Second, it minimizes similarity due to shared phonotactic preferences or overlap in phoneme inventories, which *should* be taken into account when assessing lexical distance in the present context, since similarity driven by these causes can influence the perceived similarity of words across languages.

As such, in the present study we use LDN, rather than LDND. Nevertheless, these two measures are generally strongly correlated (Holman et al., 2008a; Pompei et al., 2011; Wichmann et al., 2010), so the impact of using one over the other is likely minor.

7.5.2 Feature edit distance

7.5.2.1 Rationale for feature edit distance

As noted in our discussion of Levenshtein distance in the paper, a notable issue with this measure is that it treats all character transformations as equal, even though this does not accurately represent differences in distances as perceived by learners. For example, this means that the English word “fish” /fɪʃ/ has an equal and maximal LDN of 1 from both the corresponding Spanish word (“pez” /pes/) and the Hebrew one (“דג” /dag/), even though the English word is closer phonologically and etymologically to the Spanish word than to the Hebrew one, and could be considered a cognate of the first but not the second.

A potential way to mitigate this issue is to assign different weights to different character transformations, based on the phonological features of the associated segmental units. The resulting measure, which can be viewed as a modified form of Levenshtein distance, is referred to as *phonological edit distance*, *feature edit distance*, or *feature distance* (FD)(Allen & Becker, 2015; Eden, 2018; Fontan et al., 2016; K. C. Hall et al., 2017; Kondrak, 2000; Manurung et al., 2008; McCoy & Frank, 2018; Mortensen et al., 2016; Sanders & Chin, 2009; Schepens, Dijkstra, et al., 2013; L. Zhang, 2018). For example, when using FD, substituting /ʃ/ with /z/ would generally incur a lower penalty than substituting it with /g/, since /ʃ/ and /z/ are share the same value on more phonological features, such as being coronal, so they can be considered more similar to each other from a phonological perspective.

7.5.2.2 Limitations of feature edit distance

Though FD might be able to capture phonological similarity more accurate than LD, we decided to use LD(N) as the key measure of similarity in our study, for two main reasons.

First, while there is extensive validation for the use of LD based on research in several fields (as shown under "Validation of Levenshtein distance" in this supplementary information), there is little validation of FD in similar contexts. As such, while LD might potentially be less linguistically motivated than FD, we do know based on prior research that it is able to predict linguistics outcomes fairly well—including when used to predict the influence of crosslinguistic similarity on L2—whereas we do not yet know the same for FD. In fact, the limited research that did investigate the use of FD and similar measures did not find that they are necessarily better predictors of linguistic outcomes than simple Levenshtein distance (Wieling et al., 2007; Wieling, Nerbonne, et al., 2014). For example, as Wieling et al. (2007, p. 93) state:

> It was found that generally speaking the binary versions approximate perceptual distances better than the feature- based and acoustic-based versions. The fact that segments differ appears to be more important in the perception of speakers than the degree to which segments differ. Therefore we will use the binary version of Levenshtein distance in this article…

Second, the simplicity of LD (compared to FD) presents advantage for replication of analyses, the generalizability of findings, the comparison of findings across studies, and the minimization of researcher degrees of freedom. Specifically, while LD is generally implemented in consistent manner across the various software packages that offer it, which means that calculating LD using different packages/software will lead to the same results, this is not the case of FD, which depends heavily on factors such as:

- Which phonological features are taken into account (Gooskens & Heeringa, 2004; Nerbonne & Heeringa, 1997).[116]
- What weights should be assigned to differences in feature values, and how substitutions should be weighted compared to insertions/deletions.
- Whether different weights should be assigned to different features, and if so, then what weights. This is compounded by the fact that different features could potentially be weighted differently for different populations (e.g., speakers of different L1s, who perceive the different features differently) and in different contexts (e.g., when it comes to assessing perceived distance vs. intelligibility).
- How this distance should be normalized.[117]

Furthermore, in this regard, there is also the question of whether to use FD in particular, or a similar measures that attempts to capture crosslinguistic similarities, such as *pointwise mutual information* (PMI)(Wieling, Bloem, et al., 2014) or *naive discriminative learning* (NDL)(Wieling, Nerbonne, et al., 2014).

In summary, although FD might be more linguistically motivated than LD, it is not clear that this is the case and that FD is a better predictor of linguistic outcomes. Furthermore, much methodological work needs to be done on FD to validate and standardize its use, before it can be used with confidence by researchers.

[116] For example, in the case of the *PanPhon*, which we use in the present research, /a/ and /æ/ have an FD of 0, since the two segments share identical values for all features in the package's dataset. This can lead to situations where two entries have an FD = 0 but an LD > 0, as in the case of /bambu/ and /bæmbu/ ('bamboo'), which have an FD = 0 but an LD = 1.

[117] Generally, FD is normalized into FDN in a similar manner as LD, though its theoretical maximum is based on the number of segmental units in the longer string, rather than on the number of characters (e.g. /t͡sː/ is viewed as a single segmental unit). However, unlike in the case of LD, where substitutions of non-identical characters always incur a cost of 1, in the case of FD substitutions of non-identical characters generally (depending on the specific type of FD) incur a cost <1, so the theoretical maximum is not actually the length of the longer string, which raises questions regarding what the theoretical maximum should be. For example, should it be based on the maximal number of insertions/deletions together with the maximal possible theoretical substitution given the *global* set of segmental units, or the maximal possible substitution given the *local* set of segmental units? Furthermore, since nearly all substitutions are going to incur a lower cost than the maximum, is it even appropriate to use this maximum in the normalization process?

7.5.2.3 Our technical approach

We built models that use FD as a predictor, to supplement our main models (which use LD). However, these models, should be interpreted with caution, given the limitations of FD that we discussed above.

To calculate FD for our models, we used *PanPhon*, a Python package that relates IPA segments—both simple (e.g. /t/) and complex (e.g. /t͡sː/)—to their definitions in terms of articulatory features (Mortensen et al., 2016).[118] This includes the following 22 features:

syl [±syllabic]. Is the segment the nucleus of a syllable?

son [±sonorant]. Is the segment produced with a relatively unobstructed vocal tract?

cons [±consonantal]. Is the segment consonantal (not a vowel or glide, or laryngeal consonant)?

cont [±continuant]. Is the segment produced with continuous oral airflow?

delrel [±delayed release]. Is the segment an affricate?

lat [±lateral]. Is the segment produced with a lateral constriction?

nas [±nasal]. Is the segment produced with nasal airflow?

strid [±strident]. Is the segment produced with noisy friction?

voi [±voice]. Are the vocal folds vibrating during the production of the segment?

sg [±spread glottis]. Are the vocal folds abducted during the production of the segment?

cg [±constricted glottis]. Are the vocal folds adducted during the production of the segment?

ant [±anterior]. Is a constriction made in the front of the vocal tract?

cor [±coronal]. Is the tip or blade of the tongue used to make a constriction?

distr [±distributed]. Is a coronal constriction distributed laterally?

lab [±labial]. Does the segment involve constrictions with or of the lips?

hi [±high]. Is the segment produced with the tongue body raised?

[118] The information here is based on version 0.18 of PanPhon. Note that there are also other tools for calculating such distance, such as the *abydos* library in Python (Little, 2018); we chose PanPhon for its features and documentation, but more extensive testing and validation is needed to compare different packages.

lo [±low]. Is the segment produced with the tongue body lowered?

back [±back]. Is the segment produced with the tongue body in a posterior position?

round [±round]. Is the segment produced with the lips rounded?

velaric [±velaric]. Is the segment produced using a velaric airstream mechanism?

tense [±tense]. Is the segment produced with an advanced tongue root?

long [±long]. Does the segment take up two units of length?

This list is based on the information in Mortensen et al. (2016, p. 3478), Mortensen (2015), and Mortensen (personal communication, December 6, 2019)

PanPhon contains the data on these 22 features for different IPA segment, with 3 possible values: in cases where the feature value is specified, it is marked as either + or -, and in cases where it is unspecified, it is marked as 0. For example, table 41 contains the sample values for some of the characters in the PanPhon database.

Table 41. A sample of characters from the PanPhon database, which is used to calculate FD.

ipa	syl	son	cns	cnt	dlr	lat	nas	str	voi	sg	cg	ant	cor	dst	lab	hi	lo	bac	rnd	vel	tns	lng
p	-	-	+	-	-	-	-	0	-	-	-	+	-	0	+	-	-	-	-	-	0	-
z	-	-	+	+	-	-	-	0	+	-	-	+	+	-	-	-	-	-	-	-	0	-
ɲ	-	+	+	-	-	-	+	0	+	-	-	-	-	0	-	+	-	-	-	-	0	-
ɡ	-	-	+	-	-	-	-	0	+	-	-	-	-	0	-	+	-	+	-	-	0	-
ɡːʲ	-	-	+	-	-	-	-	0	+	-	-	-	-	0	-	+	-	-	-	-	0	+
t͡ʃ	-	-	+	-	+	-	-	0	-	-	-	-	+	+	-	-	-	-	-	-	0	-
u	+	+	-	+	-	-	-	0	+	-	-	0	-	0	+	+	-	+	+	-	+	-
ɑ	+	+	-	+	0	-	-	0	+	-	-	0	-	0	-	-	+	+	-	-	+	-
ɑː	+	+	-	+	0	-	-	0	+	-	-	0	-	0	-	-	+	+	-	-	+	+

Note. The feature values are taken directly from the Nov 11, 2019 release of PanPhon. Some feature names here are trimmed here due to space constraints.

Specifically, we used the *partial_hamming_feature_edit_distance* function,[119] which calculates FD in the following manner:

- An edit that involves an insertion or a deletion incurs a cost of 1.
- An edit that involves going from a certain feature value to an *opposite* feature value incurs a cost of 1/22. For example, if a segment that is [+back] is substituted with a segment that is [-back], a cost of 1/22 is incurred for that particular feature edit.
- An edit that involves going from a *specified* feature value to an *unspecified* feature value and vice versa incurs a cost of 1/44. For example, if a segment that is [+back] is substituted with a segment whose [back] feature is unspecified, a cost of 1/44 is incurred for that particular feature edit.
- An edit that involves going from a certain feature value to an *identical* feature value incurs no cost. For example, if a segment that is [+back] is substituted with a segment that is also [+back], no cost is incurred for that particular feature edit.

The resulting FD was normalized into FDN by dividing it by the length of the longer string in the pair, based on the number of segmental units (e.g., /t͡sː/), since FD focuses on segmental units rather than characters.

Note that whereas LD is standardized, FD is not, as mentioned in the previous section. As such, the FD that we calculated here should be viewed as only one type of FD, and other types of FD are calculated differently and may lead to different outcomes.

7.5.2.4 Descriptive statistics for FDN values

There was a moderate-to-strong correlation between FDN and LDN in both the Swadesh lists (*r* = .40, *95% CI* = [.28, .50], *p* < .001) and the parallel dictionaries (*r* = .47, *95% CI* = [.45, .49], *p* < .001). This suggests that although these two measures have a strong association, as can be expected, they capture substantially different aspects of crosslinguistic distance, and the

[119] Alternative functions are available for this purpose in PanPhon. We selected this function because it offered a balance between the two other main functions: *feature_edit_distance*, where insertion/deletions are treated the same as substitutions, and so generally incur a cost <1 (due to the presence of unspecified features), and *hamming_feature_edit_distance*, where transformations from specified feature values to unspecified ones (and vice versa) incur a cost of 1/22, similarly to transformations to opposite feature values. It is not clear that the specific distance that we used is the best one (i.e., the one that best predicts the perceived similarity between words), which highlights the need for validation and standardization of this measure. Nevertheless, this is not crucial for the present research, as the differences between the distances that these measures lead to are small enough that they do not influence our findings.

use of one rather than the other might influence the results of analyses, at least to some degree.[120]

Figure 39 and Table 42 contain information about the FDN between the L1s in the sample and English. The FDN of all word pairs is available in the data files in the OSF repository (under “Lexical distance & frequency data”).

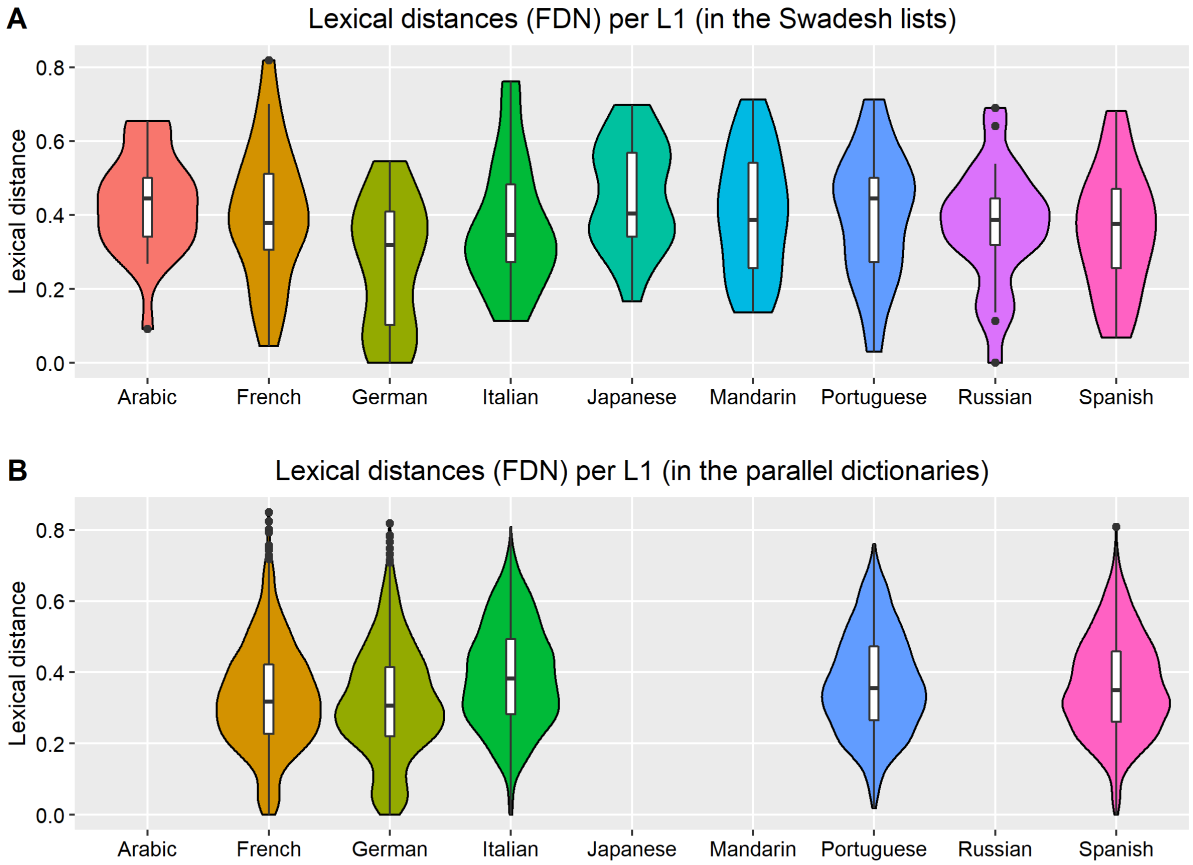


Figure 39. Lexical distance between L1 words English, per L1 in each dataset. The distance is equal to the phonological *FDN* between L1 words and their most lexically similar English counterpart. Within the boxplots, the line inside the box indicates the median, the lower and upper hinges indicate the 1st and 3rd quartiles, the whiskers indicate 1.5 interquartile ranges (IQR) past the hinges, and the dots indicate outliers beyond that. The violin plots indicate an estimate of the probability density of lexical distance for each L1, which can be viewed as the likelihood that a word in each L1 will have a certain lexical distance, where increased width indicates greater likelihood. Data is based on 25 words per L1 in the Swadesh lists and 1,103 words per L1 in the parallel dictionaries (i.e., after the removal of multi-word entries).

[120] Although we do not expect it to change the null findings in the present study, both because past studies found an effect of crosslinguistic similarity while using LD(N), and because the correlation between LDN and FDN means that we would expect to find at least some effect of similarity in our sample, which is not the case.

Table 42. Statistics about the lexical distances (FDN) between the L1s and English in each dataset. L1s are arranged in order of increasing mean lexical distance in the Swadesh lists.

	Swadesh lists					Parallel dictionaries				
L1	mean	SD	median	IQR	range	mean	SD	median	IQR	range
German	.271	.17	0.32	0.10-0.41	0.00-0.55	.316	.15	0.31	0.22-0.42	0.00-0.82
Spanish	.357	.16	0.38	0.26-0.47	0.07-0.68	.362	.14	0.35	0.26-0.46	0.00-0.81
Russian	.370	.16	0.39	0.32-0.45	0.00-0.69	-	-	-	-	-
Italian	.371	.17	0.35	0.27-0.48	0.11-0.76	.390	.14	0.38	0.28-0.49	0.00-0.81
Portuguese	.400	.17	0.45	0.27-0.50	0.03-0.71	.370	.14	0.36	0.27-0.47	0.02-0.76
French	.404	.19	0.38	0.31-0.51	0.05-0.82	.330	.15	0.32	0.23-0.42	0.00-0.85
Mandarin	.405	.17	0.39	0.26-0.54	0.14-0.71	-	-	-	-	-
Arabic	.432	.14	0.45	0.34-0.50	0.09-0.65	-	-	-	-	-
Japanese	.455	.14	0.40	0.34-0.57	0.17-0.70	-	-	-	-	-

Note. The distance here is the phonological *FDN* from the closest synonym, calculated for the single-word entries in each dataset. There were 225 entries in the Swadesh lists (i.e., rows with an English word and all its corresponding counterparts in a certain L1), with 25 entries for each of the 9 L1s in the dataset. There were 5,515 entries in the parallel dictionaries, with 1,103 for each of the 5 L1s. All counts are after the removal of multi-word entries.

Several key observations can be made about these distances.

First, FDN is much more evenly distributed within each L1 than LDN, primarily due to the lack of ceiling effect present in LDN (i.e., the tendency of words to have the maximal possible LDN of 1). This can likely facilitate analyses using this distance, but it does not necessarily more accurately represent distance between words as perceived by learners.

Second, there are some similarities and differences in the per-L1 differences here compared to those based on LDN, as shown in Table 43 below. Specifically, the similarities are that German is ranked as the closest L1 to English, and that all the Romance L1s (French, Italian, Spanish, and Portuguese) are ranked as closer than all the non-Indo-European L1s (Arabic, Japanese, and Mandarin). The differences are that the ranking is different within the Romance L1, the Indo-European L1s, and the non-Indo-European L1s, and that there are also several differences across these groups, including, most notably, that in FDN Russian is ranked as substantially closer to English than Portuguese and Mandarin, and that French is ranked as being practically as distant from English as Mandarin. These distances are not directly reflective of those between the languages, since they include only single-word entries (as discussed in more detail under the "Validation of Levenshtein distance" in the supplementary information). Nevertheless, these as shown in the aforementioned section, these distances are expected to be close to the "real" distances between these languages, and as such the results for FDN are highly unexpected, especially in the case of French. This suggests that the present FDN measure is not better than LDN at quantifying crosslingusitic distance.

Table 43. Comparison of the ranking of the L1s based on their distance from English in the Swadesh lists, separately for LDN and FDN.

	LDN		FDN	
Rank	L1	Mean	L1	Mean
1	German	.622	German	.271
2	Italian	.776	Spanish	.357
3	Spanish	.808	Russian	.370
4	French	.813	Italian	.371
5	Portuguese	.848	Portuguese	.400
6	Japanese	.864	French	.404
7	Russian	.881	Mandarin	.405
8	Arabic	.887	Arabic	.432
9	Mandarin	.924	Japanese	.455

7.5.2.5 FDN-based models

As with our main models, we used the normalized version of this distance (FDN), which we scaled (by multiplying it by 10) and centered.

We initially built these models using the same fixed and random effects as in our main models. However, the Swadesh-based models in both subcorpora had issues with singular convergence (due to the intercepts and slopes of the L1 random effect), and the parallel-based models did not converge at all.[121]

As such, below (in Tables 44 and 45) we present the results for FDN-based models without the L1 random effect. However, this does not substantially influence our findings, since this effect was very weak in the FDN-based models that contained it and did converge, and the results of the models were functionally identical regardless of the inclusion of this effect, as was the case for the LDN-based models (see the “Models without the L1 random effect” in the supplementary information).

These tables show that the FDN-based models replicate the key findings of the LDN-based models, with a similar null effect of distance and its interaction with proficiency (*B* = 0.00–0.01, corresponding to *IRR* = 1.00–1.01), together with strong task effects.

[121] Specifically, they had a “gradient function must return a numeric vector of length 7” error and a “NA/NaN function evaluation” warning.

In addition, we also built FDN-based models using only data from German speakers. This is both to replicate the associated LDN-based models, and because the FDN-based results for the German speakers were consistent with the LDN-based results and with what is expected based on general language classification (as discussed under "Validation of Levenshtein distance" in the supplementary information), while also being the L1 that is closest to English.

The results of these models are shown in Tables 46 and 47. As with the German-based models that used LDN as the measure of distance, these models replicate the key findings of the main models, in terms of the lack of a substantial effect of distance or of its interaction with proficiency, and in terms of the strong task effects.[122]

Overall, the results from the FDN-based models complement those of the LDN-based main models, and suggest that the null effect in the main models should not be attributed to LDN failing to fully capture the phonological overlap between words, something that is also supported by past validation of Levenshtein distance. However, given the limitations of FD that were above, both in general and within this sample, more work on validating and standardizing FD and similar measures is needed before a conclusive statement can be made on the influence of its use in this context.

[122] The one notable difference is the much weaker effect of frequency here for the parallel-based model in the first corpus, together with an associated increase in the magnitude of the intercept, as is the case with the corresponding LDN-based model. We do not have a clear explanation for this, but it is not crucial for the present analyses, given that the key findings replicate despite of this, and that this was an issue for only one of the four models.

Table 44. Results of the mixed-models with *FDN* as the distance measure, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological FDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-10.32	0.16	0.00	<0.01	-65.27	<.001	-9.85	0.14	0.00	<0.01	-68.84	<.001
Distance	0.01	<0.01	1.01	<0.01	3.60	<.001	0.00	<0.01	1.00	<0.01	-0.13	.898
Proficiency	-0.04	0.02	0.96	0.02	-2.10	.035	0.00	0.02	1.00	0.02	-0.24	.813
Frequency	3.29	0.21	26.94	5.67	15.65	<.001	3.15	0.19	23.29	4.44	16.52	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	1.03	.302	0.00	<0.01	1.00	<0.01	1.45	.148
Random effects												
Learner_τ_{00}	0.07						0.24					
Task_τ_{00}	0.40						0.33					
Word_τ_{00}	0.38						0.46					
Task:Word_τ_{00}	1.84						1.36					

Table 45. Results of the mixed-models with *FDN* as the distance measure, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological FDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-12.84	0.06	0.00	<0.01	-210.04	<.001	-12.58	0.05	0.00	<0.01	-246.81	<.001
Distance	0.01	<0.01	1.01	<0.01	2.60	.009	0.00	<0.01	1.00	<0.01	0.61	.542
Proficiency	0.12	0.01	1.13	0.01	10.14	<.001	0.04	0.01	1.04	0.01	4.22	<.001
Frequency	2.90	0.06	18.16	1.05	49.95	<.001	2.97	0.05	19.50	0.99	58.51	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	1.91	.056	0.00	<0.01	1.00	<0.01	0.67	.501
Random effects												
Learner_τ_{00}	0.03						0.05					
Task_τ_{00}	0.03						0.11					
Word_τ_{00}	0.45						0.65					
Task:Word_τ_{00}	2.32						1.50					

Table 46. Results of the mixed-models with *FDN* as the distance measure, for the Swadesh-based samples, using only data from German speakers. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological FDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-9.81	0.21	0.00	<0.01	-47.29	<.001	-9.35	0.17	0.00	<0.01	-55.11	<.001
Distance	0.06	0.10	1.07	0.11	0.63	.528	0.01	0.08	1.01	0.08	0.15	.884
Proficiency	-0.07	0.02	0.93	0.02	-3.13	.002	-0.01	0.02	0.99	0.02	-0.55	.579
Frequency	2.69	0.24	14.79	3.50	11.38	<.001	2.69	0.19	14.72	2.79	14.19	<.001
Dist:Prof	0.03	0.01	1.03	0.01	2.31	.021	0.00	0.01	1.00	0.01	0.46	.648
Random effects												
Learner_τ_{00}	0.00						0.27					
Task_τ_{00}	0.13						0.20					
Word_τ_{00}	0.21						0.31					
Task:Word_τ_{00}	1.80						1.15					

Table 47. Results of the mixed-models with *FDN* as the distance measure, for the parallel-based samples, using only data from German speakers. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological FDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-15.47	0.07	0.00	<0.01	-237.29	<.001	-12.46	0.06	0.00	<0.01	-205.98	<.001
Distance	0.00	0.03	1.00	0.03	-0.10	.922	-0.02	0.03	0.98	0.02	-0.99	.322
Proficiency	-0.03	0.01	0.97	0.01	-3.41	.001	0.02	0.01	1.02	0.01	2.77	.006
Frequency	0.11	0.06	1.12	0.07	1.75	.080	2.65	0.06	14.19	0.85	44.37	<.001
Dist:Prof	0.02	<0.01	1.02	<0.01	5.43	<.001	0.01	<0.01	1.01	<0.01	3.85	<.001
Random effects												
Learner_τ_{00}	0.00						0.03					
Task_τ_{00}	0.03						0.03					
Word_τ_{00}	0.35						0.38					
Task:Word_τ_{00}	2.30						1.66					

7.5.3 Additional descriptive information

7.5.3.1 Correlations of distance, frequency, and word use

Figure 40 contains basic scatterplots with the usage of the target English words in relation to their lexical distance from the corresponding L1 words. These plots show that the datasets contain words with a broad range of lexical distances, and a broad range of rates of usage. In addition, there appears to be a weak positive association between lexical distance and word usage, since the words with the higher rates of usage are almost exclusively located on the right. This is contrary to the negative correlation that we expect, whereby higher distance is associated with reduced usage. However, this could be due to confounds such as the baseline frequency of the English words, which our mixed-models address.

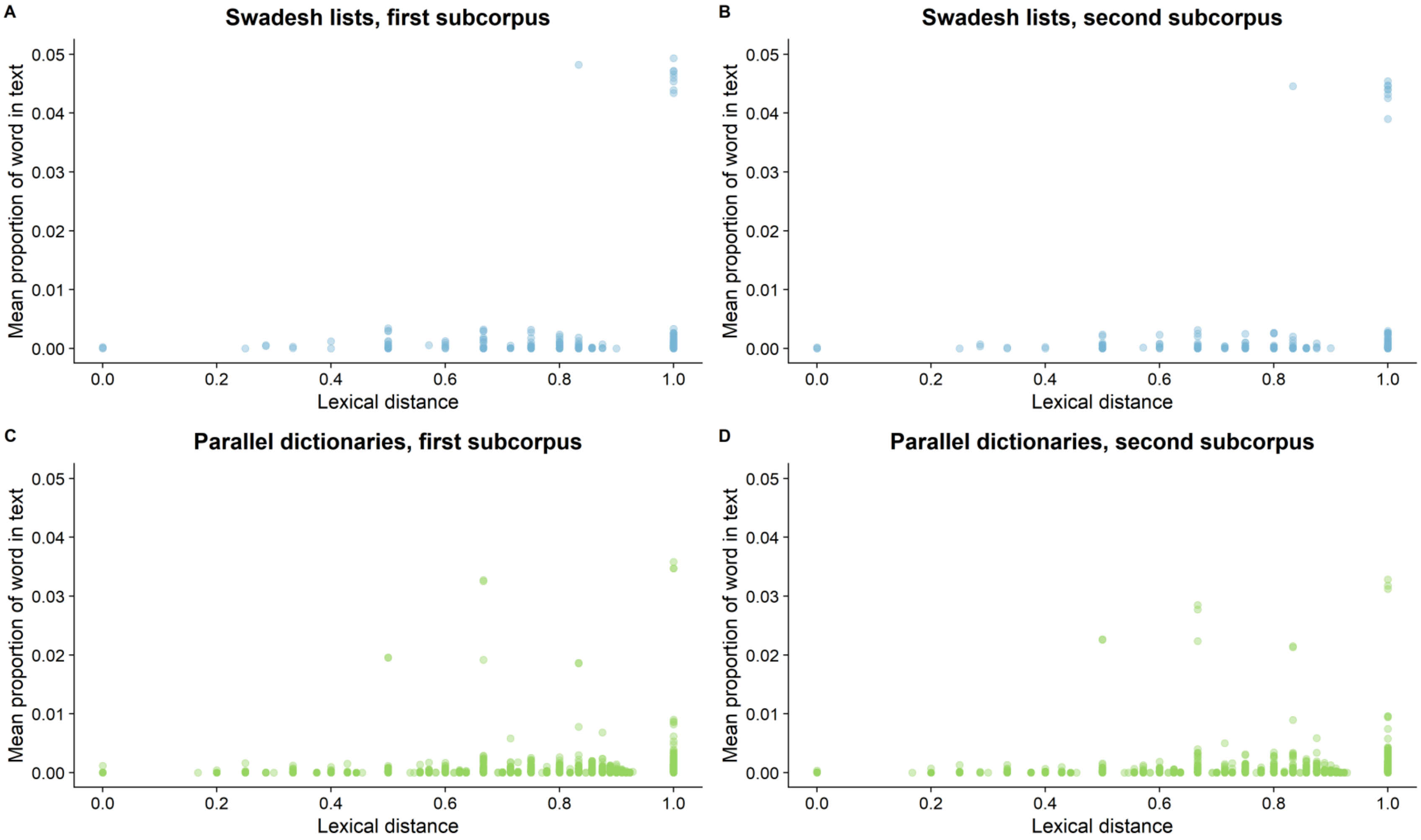


Figure 40. Usage of the target English words, based on their mean proportion in texts (where the proportion of a word in each text is based on the number of times it is used there, divided by the total number of words in the text). Each point is a combination of a target word and a specific L1, since the different L1s can have different distances from English for any given word. Darker shading indicates an overlap in points.

Table 48 shows the raw correlations between lexical distance, baseline frequency of the words in English, and the rate of usage of the L2 English words in the present sample.

For both lexical-distance datasets, there is a significant and substantial positive correlation between the baseline frequency of words and their rate of use in the learner sample, though this correlation is stronger for the words in the Swadesh lists (*r* = .39–.41) than in the parallel dictionaries (*r* = .17–.18). In addition, in the Swadesh lists there is also a significant and substantial positive correlation (*r* = .18) between the lexical distance of words and their frequency, meaning that more distant words are more frequent, but this correlation is not substantial (*r* = .03) in the parallel dictionaries.

In addition, there is a weak positive correlation between distance and usage for the Swadesh-based samples (*r* = .10–.11), which might be attributable to the distance-frequency and frequency-usage correlations. This is opposite to the association that we would expect between distance and usage if there was a cognate facilitation effect (assuming no other factors played a role), since *decreased* crosslinguistic distance (i.e., *increased* similarity/cognancy) should lead to *increased* word use. In the case of the parallel dictionaries, there is functionally no correlation between distance and usage (*r* = .01), which is expected given the almost null correlation between distance and frequency in this dataset, together with the weaker correlation between frequency and word use.

The difference in correlations between the Swadesh lists and the parallel dictionaries can be attributed, in part, to the fact that the parallel dictionaries contain a broader range of words in terms of their baseline English frequencies, including ones that are lower-frequency than in the Swadesh lists (Zipf frequency range of 1.87–7.41 in the parallel dictionaries, compared to 4.15–7.11 in the Swadesh lists). However, as shown in Table 49, when this difference is largely eliminated, by selecting a subset of the parallel dictionaries containing only words with a Zipf frequency of 4.15 and above (as in the Swadesh lists),[123] the *distance-frequency* and *frequency-usage* correlations increase but remain weaker than in the Swadesh lists (respectively, *r* = .07 and *r* = .23–.25), and the *distance-usage* correlation remains functionally zero (*r* = .01 in both corpora).

[123] Though there is still a small but significant difference (*W* = 457718, *p* < .001) in the mean frequency of words between the datasets, where the mean Zipf frequency in the Swadesh lists is 5.24 (*SD* = 0.72), and the mean frequency in the parallel dictionaries is 4.91 (SD = 0.54).

Table 48. The raw correlations between lexical distance, frequency, and word usage, presented in the form of *Pearson's r [95% CI] (p)*.

	Swadesh lists		Parallel dictionaries	
	First corpus	Second corpus	First corpus	Second corpus
Distance-Frequency [a]	.18 [.05, .30] (.007)		.03 [.01, .06] (.013)	
Frequency-Usage [b]	.39 [.38, .39] (<.001)	.41 [.40, .41] (<.001)	.17 [.16, .17] (<.001)	.18 [.18, .18] (<.001)
Distance-Usage [b]	.10 [.10, .11] (<.001)	.11 [.10, .11] (<.001)	.01 [.01, .01] (<.001)	.01 [.01, .01] (<.001)

[a] The distance-frequency correlation depends only on the source of lexical-distance data (i.e., it is corpus-*independent*). *N* = 225 for the Swadesh lists (based on 25 entries for each of the 9 L1s included there), and *N* = 5,515 for the parallel dictionaries (based on 1,103 entries for each of the 5 L1s included there).

[a] *Usage* is based on the mean proportion of words in each text (based on the number of times each word is used there, divided by the total number of words in the text). As such, this measure is corpus-*dependent*. Sample sizes for it were 212,500 (Swadesh, first), 159,750 (Swadesh, second), 5,235,941 (parallel, first), and 3,915,650 (parallel, second); this is based on the number of lexical-distance entries multiplied by the number of available texts.

Table 49. The raw correlations between lexical distance, frequency, and word usage, presented in the form of *Pearson's r [95% CI] (p)*. Data is based on words in the parallel dictionaries with a Zipf frequency ≥ 4.15 (*n* = 3195, 57.93% of the total words in the parallel dictionaries). For the corpus-dependent correlations (i.e., those involving usage), there were 3,033,333 observations in the first corpus, and 2,268,450 in the second.

	First corpus	Second corpus
Distance-Frequency	.07 [.04, .10] (<.001)	
Frequency-Usage	.23 [.22, .23] (<.001)	.25 [.25, .25] (<.001)
Distance-Usage	.01 [.01, .01] (<.001)	.01 [.01, .01] (<.001)

One possibility that was raised, based on the findings of the mixed-models in the paper, is that the cognate facilitation effect does not exist, and was found in other studies due to the confounding influence of factors such as frequency, which we controlled for in the models. While this would be a novel finding in its own right, we do not believe that this is the case.

This is because past studies have found evidence of the cognate facilitation effect even when frequency is controlled for, so we would expect to find this effect here too (Bosma et al., 2019; Carrasco-Ortiz et al., 2021; Casaponsa et al., 2015; Costa et al., 2000; De Wilde et al., 2020, 2021; Hoshino & Kroll, 2008; Otwinowska et al., 2020; Otwinowska & Szewczyk, 2019; Poort & Rodd, 2017; Sadat et al., 2016; Sheng et al., 2016; van de Ven et al., 2019; J. Zhang et al., 2019; Y. Zhu & Mok, 2020). Similarly, in the case of task effects, the aforementioned studies found cognate facilitation using a wide range of methods, including ones where task effects, as conceptualized in the present study, do not play a role, since they were focused primarily on experiment-based investigation of language processing, so it does not appear the us controlling for task effects could explain the lack of cognate facilitation either.

In addition, the correlations that we found here do not lead to a cognate facilitation effect, even without controlling for proper background factors. Specifically, in the case of the Swadesh lists, based on the positive distance-frequency and frequency-usage correlations, we would expect to find an effect *opposite* to cognate facilitation, in the sense that *increased* distance (i.e., *reduced* similarity) will correlate with increased word use, which is in fact what we find for the distance-usage correlation. Furthermore, in the case of the parallel dictionaries, we would *not* expect to find a similar effect at all, since the correlation between distance and frequency is functionally zero.

Overall, the extensive evidence from past studies shows that the cognate facilitation effect exists even when frequency and other factors are controlled for. Furthermore, the raw correlations between the key variables in our study (lexical distance, baseline frequency, and L2 word usage) show that, when background factors are not properly controlled for, we would expect to find either a null effect or an opposite effect than cognate facilitation. As such, the absence of the cognate facilitation effect in our main models is a novel theoretical finding, that is not merely attributable to the fact that we control for background factors such as frequency.

7.5.3.2 Frequency-ratio descriptive statistics

Table 50 contains descriptive statistics regarding the frequency ratio of the words in the samples, as visualized in Figure 3 of the paper (in the beginning of the Results section). It shows that, on average, target English words were used in equal rates in the sample as in baseline English (i.e., had a frequency ratio near 1). However, all samples contained a range of words with different frequency ratios (total range 0.70–1.58), and this rate was greater in the parallel-based samples, likely due to the inclusion of very low-frequency words. In addition, this inclusion is likely also the reason why more of the words from the parallel dictionaries did not appear in the parallel-based samples at all, as indicated by the substantially higher rate of words with a frequency of 0 in the parallel dictionaries.

Table 50. Descriptive statistics regarding *frequency ratio*, which is the frequency of a word in a given sample divided by its baseline frequency in English. The baseline frequency in English is based on the same frequency measure that we use throughout the paper, as discussed in the "Baseline word frequency" section of the paper. The frequency of use per sample is calculated separately for each combination of a target word and a specific L1, since different L1s can have different distances from English for any given word. The frequencies within the sample are based on 8,500 texts for the Swadesh lists in the first subcorpus, 6,390 for the Swadesh lists in the second subcorpus, 4,747 for the parallel dictionaries in the first subcorpus, and 3,550 for the parallel dictionaries in the second subcorpus. This corresponds to 212,500 observations (number of words per L1 times the number of texts in the sample) for the Swadesh lists for the first subcorpus, 159,750 for the Swadesh lists in the second subcorpus, 5,235,941 for the parallel dictionaries in the first subcorpus, and 3,915,650 for the parallel dictionaries in the second subcorpus.

			Frequency of 0 [b]		Frequency ratio [c]			
Distance dataset	Subcorpus	Words [a]	*n*	%	*mean*	*median*	*SD*	*range*
Swadesh lists	first	225	31	13.78	1.03	1.03	0.09	0.70–1.19
Swadesh lists	second	225	31	13.78	1.00	1.00	0.07	0.76–1.19
Parallel dictionaries	first	5,515	2,691	48.79	1.02	1.01	0.12	0.71–1.57
Parallel dictionaries	second	5,515	2,818	51.10	1.01	1.00	0.10	0.70–1.58

[a] *Words* is equal to the number of L1s in the distance dataset (9 in the Swadesh lists, 5 in the parallel dictionaries), times the number of words per L1 (25 in the Swadesh lists, 1,103 in the parallel dictionaries).

[b] Words that did not appear in the sample were assigned a Zipf frequency of 0, in line with Speer (2020), and consequently have a frequency ratio of 0 here. *n* represents the number of such words in the sample, and the *%* represents the percent of such words out of the total words in the sample.

[c] All the *frequency ratio* statistics were calculated while excluding cases with a frequency of zero. A ratio of 1 indicates that a word is used in an equal rate in the sample and in baseline English, whereas a ratio >1 indicates that the word is used more frequently in the sample, and a ratio <1 indicates the opposite.

7.5.4 Incidence rate ratio (IRR)

As noted in the body of the paper, we exponentiated the coefficient estimates in the mixed-models to derive an *incidence rate ratio* (IRR), in order to facilitate the interpretation of the results, and the *standard errors* (SEs) of the coefficients were then scaled by multiplying them by the exponentiated coefficient estimates (Hox et al., 2018; Sedgwick, 2010).

The IRR itself can be interpreted as the expected change in the rate of the response variable as a factor of a 1-unit increase in the predictor. For example, an IRR of 2 means that a 1-unit increase in the predictor doubles the rate of response (i.e., doubles the rate of use of the target word), while an IRR of 0.5 means that a 1-unit increase in the predictor halves it. An IRR of 1 corresponds to a coefficient estimate (*B*) of 0, as there is no expected change in the response variable as a result of a change in the predictor.

It is important to note that when combining multiple coefficients, you should *not* add the exponentiated coefficients, but rather multiply them, which is equivalent to exponentiating the added coefficients. For example, consider a situation where you are predicting the IRR of a word that is 1 unit more frequent than some baseline level, in a learner whose proficiency is 1 unit higher than some baseline level. If the raw coefficient of frequency is 0.5 and that of proficiency is 0.3, then the IRR will be:

$$e^{0.5} \times e^{0.3} = e^{(0.5+0.3)} = 2.2$$

In addition, if you want to predict the IRR of a word that is 1 unit *less* frequent, then you need to take the inverse of the IRR of a word that is 1 unit *more* frequent, since this is equivalent to exponentiating the negative of the associated coefficient. For example, if the coefficient is 0.5, then the IRR of a word that is 1 unit *less* frequent than some baseline level is:

$$e^{-0.5} = \frac{1}{e^{0.5}} = 0.6$$

7.5.5 Random effects

7.5.5.1 Random slopes

Initially, we tested several potential mixed-effects models, with random slopes of *lexical distance* for the *learner*, *L1*, *task*, and *word* random effects (separately for each one). For the models based on the parallel dictionaries, only the model with random slopes for *L1* converged properly, as the other models either had problems with singular convergence or did not converge at all, even though they were tested on their own (i.e., as a single random slope, before combining multiple ones).

Given this, and given that the goal was to use a consistent random-effects structure across all models, we included only random slopes of *distance* for *L1* in these models. However, as shown in the results section of the main paper, this does not appear to be an issue given our particular findings, since the main concern with omitting random slopes is an increased rate of Type I error (Matuschek et al., 2017; Winter, 2019), but our key findings provide support for the null hypothesis.

7.5.5.2 Random intercepts by text

We considered adding to the models a random effect (random intercepts) for each *text* in the sample. However, there is substantial overlap between this and the *learner* random effect, since, as noted in the paper, most learners only had a single text in the sample.[124] In addition, we also had the *task* random effect, which accounts for further variance that may be associated with specific texts (each learner had only a single text per task).

When we attempted to build models that included the *text* random effect in addition to *learner*, in the case of the parallel-based models, the model did not converge for the first corpus, and had convergence warning for the second corpus.[125] Given this, and given that the goal was to use a consistent random-effects structure across all models, we did not include this random effect in our final models.

[124] The mean number of texts per learner was 1.36 in the first corpus and 1.41 in the second. For more details on this, see the “Sample information” document in the study’s OSF repository.

[125] In the first corpus, we had a “gradient function must return a numeric vector of length 8” error, as well as “NA/NaN function evaluation” and “restarting interrupted promise evaluation” warnings. In the second corpus, we had the same warnings as in the first corpus, but not the error.

Nevertheless, as shown Tables 51 and 52, the models that did converge with this random effect were functionally equivalent to the models without it, so excluding this effect from the main models does not make a substantial difference to our findings.[126]

[126] In addition, note that the random effect of *text* was estimated to be functionally equivalent to zero in 2 out of the 3 models that did converge, possibly because there was not sufficient information to disentangle it from the *learner* random effect.

Table 51. Results of the mixed-models with *text* as an additional random effect, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-10.32	0.16	0.00	<0.01	-65.39	<.001	-9.87	0.14	0.00	<0.01	-68.53	<.001
Distance	-0.01	0.01	0.99	0.01	-1.17	.243	-0.01	0.01	0.99	0.01	-0.38	.701
Proficiency	-0.04	0.02	0.96	0.02	-2.12	.034	0.00	0.02	1.00	0.02	-0.25	.802
Frequency	3.30	0.21	26.99	5.66	15.70	<.001	3.16	0.19	23.50	4.50	16.49	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	0.61	.543	0.00	<0.01	1.00	<0.01	-1.18	.238
Random effects												
Learner_τ_{00}	0.07						0.15					
Text_τ_{00}	0.00						0.24					
Task_τ_{00}	0.40						0.33					
Word_τ_{00}	0.38						0.46					
Task:Word_τ_{00}	1.84						1.36					
L1_τ_{00}	0.02						0.03					
L1.Distance_τ_{11}	0.01						0.03					
L1_ρ_{01}	0.55						-0.05					

Table 52. Results of the mixed-models with *text* as an additional random effect, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	*SE$_B$*	*IRR*	*SE$_{IRR}$*	*z*	*p*	*B*	*SE$_B$*	*IRR*	*SE$_{IRR}$*	*z*	*p*
(Intercept)							-12.59	0.05	0.00	<0.01	-243.41	<.001
Distance							0.01	0.01	1.01	0.01	1.04	.301
Proficiency							0.04	0.01	1.04	0.01	4.29	<.001
Frequency							2.97	0.05	19.50	0.99	58.52	<.001
Dist:Prof							0.00	<0.01	1.00	<0.01	1.09	.276
Random effects												
Learner_τ_{00}							0.04					
Text_τ_{00}							0.00					
Task_τ_{00}							0.11					
Word_τ_{00}							0.65					
Task:Word_τ_{00}							1.50					
L1_τ_{00}							0.01					
L1.Distance_τ_{11}							0.01					
L1_ρ_{01}							0.81					

7.5.5.3 Models without the L1 random effect

We built supplementary models without the L1 random effect (i.e., without random slopes of *distance* for *L1* and without random intercepts for *L1*). The main reason for this are that there was only a relatively small number of L1s in the samples, particularly in the parallel-based models, so we wanted to check whether and how removing this effect would change the estimates for the other effects.[127]

As shown Tables 53 and 54, the findings of these models mirror almost exactly those of the main models in the paper, which indicates that including or excluding the L1 random effect does not substantially influence the findings.

[127] Also, this effect was very weak and practically null. It is likely that it is underestimated to some degree due to the small number of L1s, though, given that it is very weak even in the Swadesh-based models, where there are more L1s, it is likely that it is also very weak in reality, even if slightly less so.

Table 53. Results of the mixed-models without the *L1* random effect, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-10.32	0.16	0.00	<0.01	-65.41	<.001	-9.85	0.14	0.00	<0.01	-68.76	<.001
Distance	-0.01	0.01	0.99	0.01	-2.05	.040	-0.01	0.01	0.99	0.01	-0.99	.321
Proficiency	-0.04	0.02	0.96	0.02	-2.12	.034	0.00	0.02	1.00	0.02	-0.22	.827
Frequency	3.29	0.21	26.97	5.66	15.70	<.001	3.15	0.19	23.40	4.46	16.52	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	0.72	.471	0.00	<0.01	1.00	<0.01	-1.25	.211
Random effects												
Learner_τ_{00}	0.07						0.24					
Task_τ_{00}	0.40						0.33					
Word_τ_{00}	0.38						0.46					
Task:Word_τ_{00}	1.84						1.36					

Table 54. Results of the mixed-models without the *L1* random effect, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the SD of the associated random intercepts.

	First subcorpus						Second subcorpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-12.84	0.06	0.00	<0.01	-210.05	<.001	-12.58	0.05	0.00	<0.01	-246.85	<.001
Distance	0.01	<0.01	1.01	<0.01	3.44	.001	0.01	<0.01	1.01	<0.01	1.54	.124
Proficiency	0.12	0.01	1.13	0.01	10.17	<.001	0.04	0.01	1.04	0.01	4.21	<.001
Frequency	2.90	0.06	18.15	1.05	49.97	<.001	2.97	0.05	19.48	0.99	58.51	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	1.21	.226	0.00	<0.01	1.00	<0.01	1.07	.283
Random effects												
Learner_τ_{00}	0.03						0.05					
Task_τ_{00}	0.03						0.11					
Word_τ_{00}	0.44						0.65					
Task:Word_τ_{00}	2.32						1.50					

7.5.6 Model diagnostics (assumption checks)

7.5.6.1 Residual plots

7.5.6.1.1 Rationale for diagnostic approach

When interpreting the diagnostic plots, we follow two notable recommendations from Winter's (2019) relevant work, and namely the focus on visual techniques for diagnostic purposes, and the use of assumption checking on as a way to determine whether there are any major issues with the model. As Winter notes in this regard:

> Newcomers to regression modeling often find it discomforting that the assumptions are assessed visually. In fact, formal tests for checking assumptions do exist, such as the Shapiro-Wilk test of normality. However, applied statisticians generally prefer visual diagnostics (Quinn & Keough, 2002; Faraway, 2005, 2006: 14; Zuur et al., 2009, Zuur, Ieno, & Elphick, 2010). The most important reason for using graphical validation of assumptions is that it tells you more about your model and the data. [Footnote 7: Here are some other reasons: each of these tests also has assumptions (which may or may not be violated), the tests rely on hard cut-offs such as significance tests (even though adherence to assumptions is a graded notion), and the tests may commit Type I errors (false positives) or Type II errors (false negatives)…] For example, the residuals may reveal a hidden nonlinearity, which would suggest adding a nonlinear term to your model (see Chapter 8). Or the residuals may reveal extreme values that are worth inspecting in more detail. One should also remember that a model's adherence to the normality and constant variance assumptions is not a strict either/or. Faraway (2006: 14) says that 'It is virtually impossible to verify that a given model is exactly correct. The purpose of the diagnostics is more to check whether the model is not grossly wrong.'"
>
> (Winter, 2019, pp. 109-110)

The reliance on visual checks is particularly important given the large sample sizes in the present study, which can lead to statistically significant but meaningless deviations from model assumptions (Hartig, 2020).

7.5.6.1.2 Technical details

All analyses were conducted using R. The models were built using the *glmmTMB* package, which was developed for fitting generalized linear mixed models (GLMMs) (Brooks et al., 2017).[128] Analysis of residuals for the model diagnostics was performed using the DHARMa package (Hartig, 2021b). This package was chosen as it is dedicated to residual diagnostics for the type of models used in the present study (GLMMs), and it is used in the glmmTMB documentation as the package of choice for this purpose (Bolker, 2020), and it is also widely used by others for this purpose (e.g., Brooks et al., 2019; Gries, 2021).

DHARMa uses an approach to residual diagnostics that addresses common issues with such diagnostics. Full details for the package's approach to diagnostics, and for the rationale behind this approach, can be found in the package's documentation (Hartig, 2021a). However, the key points regarding this approach are the following:

> DHARMa aims at solving these problems by creating readily interpretable residuals for generalized linear (mixed) models that are standardized to values between 0 and 1, and that can be interpreted as intuitively as residuals for the linear model. This is achieved by a simulation-based approach, similar to the Bayesian p-value or the parametric bootstrap, that transforms the residuals to a standardized scale. The basic steps are:
>
> 1. Simulate new data from the fitted model for each observation.
>
> 2. For each observation, calculate the empirical cumulative density function for the simulated observations, which describes the possible values (and their probability) at the predictor combination of the observed value, assuming the fitted model is correct.
>
> 3. The residual is then defined as the value of the empirical density function at the value of the observed data, so a residual of 0 means that all simulated values are larger than the observed value, and a residual of 0.5 means half of the simulated values are larger than the observed value.
>
> …
>
> The key advantage of this definition is that the so-defined residuals always have the same, known distribution, independent of the model that is fit, if the model is correctly specified. To see this, note that, if the observed data was created from the same data-generating process that we simulate from, all values of the cumulative distribution should appear with equal probability. That means we expect the distribution of the

[128] We chose *glmmTMB* for several reasons, including that it is designed with GLMMs in mind, it supports variants of Poisson models that we used or expected to potentially need (e.g., Conway-Maxwell Poisson), it is substantially faster than many competing packages for the type of models that we built (Brooks et al., 2017), it is well-documented, it interfaces well with other relevant packages (e.g., *broom.mixed*), and it uses a similar syntax as *lme4*.

> residuals to be flat, regardless of the model structure (Poisson, binomial, random effects and so on).
>
> (Hartig, 2021a)

Specifically, for each model, we ran the four main diagnostic functions that are available in DHARMa. These are explained in detail in the DHARMa documentation (Hartig, 2021a), but we can briefly say the following regarding them and regarding their interpretation:

A. *plotQQunif*- this produces a uniform quantile-quantile plot from a DHARMa output. In a well-specified model, the residuals (black dots) should be plotted over the straight red line.

B. *plotResiduals*- this plots the residuals against rank-transformed predicted values.[129] In a well-specified model, the residuals (marked by the shaded grey background and black dots) should be spread homogeneously both vertically and horizontally, and the associated smooth spline (red dashed line) should be plotted over the mean line (solid red line at the horizontal 0.50 mark).[130] Note that, due to the large number of residuals, areas that are *shaded* darker denote a higher concentration of residuals, even if individual black points are not shown there; this is because the function uses a smooth scatterplot instead of a regular scatterplot when the number of residuals is high, to facilitate visual assessment. In addition, stars are used to mark simulation outliers (i.e., data points that are outside the range of simulated values), though it is not a judgment about the magnitude of the residual deviation.

C. *testDispersion*- this tests whether the observed data is more or less dispersed than expected under the fitted model, by comparing the variance of the observed residuals against the variance of the simulated residuals. The key outcome of this test is the ratio between the two, where a ratio < 1 indicates underdispersion, while a ratio > 1 indicates overdispersion.

D. *testZeroInflation*- this compares the observed number of zeros with the zeros expected from simulations. The key outcome of this test is the ratio between the two, where a ratio < 1 indicates that the observed data has fewer zeros than expected, while a ratio > 1 indicates that it has more zeros than expected (i.e., zero-inflation).

[129] The predicted values are rank-transformed by default, since this makes patterns easier to spot visually, especially if the distribution of predictors is skewed, as noted in the DHARMa documentation (http://web.archive.org/web/20210803085455/https://rdrr.io/cran/DHARMa/man/plotResiduals.html).

[130] Note that "a scaled residual value of 0.5 means that half of the simulated data are higher than the observed value, and half of them lower. A value of 0.99 would mean that nearly all simulated data are lower than the observed value. The minimum/maximum values for the residuals are 0 and 1." (Hartig, 2021a). Furthermore, due to the way that residuals are transformed in DHARMa, the scaled residuals in a properly fitted model are expected to have a *uniform*—rather than *normal*—distribution.

The results of the diagnostic checks for each model will be presented in their own figure in the next sub-section, in the form of a panel with 4 checks, each represented by a dedicated plot. Within each figure, plot (A) will correspond to the results from the *plotQQunif* function, plot (B) will correspond to *plotResiduals*, plot (C) will corresponds to *testDispersion*, and plot (D) will correspond to *testZeroInflation*.

Note that, as mentioned in the DHARMa documentation, some minor deviations from perfect patterns (e.g., in the residual plots) can occur due to chance, even in well-specified models. Furthermore, when assessing deviations, it is important to consider the magnitude of the deviation in addition to its significance, as even negligible deviations can be significant in large samples.

7.5.6.1.3 Diagnostic plots

The diagnostic plots for the Swadesh-lists models appear in Figures 41 and 42. In each figure, (A) contains the QQ plot, (B) contains the residual plot, (C) contains the dispersion test, and (D) contains the zero-inflation test. These diagnostic checks suggest that the models are fairly well-specified, though they have some underdispersion, particularly in the first corpus; the potential consequences of this are discussed at the end of this sub-section, after the diagnostic plots for the parallel-dictionaries models.

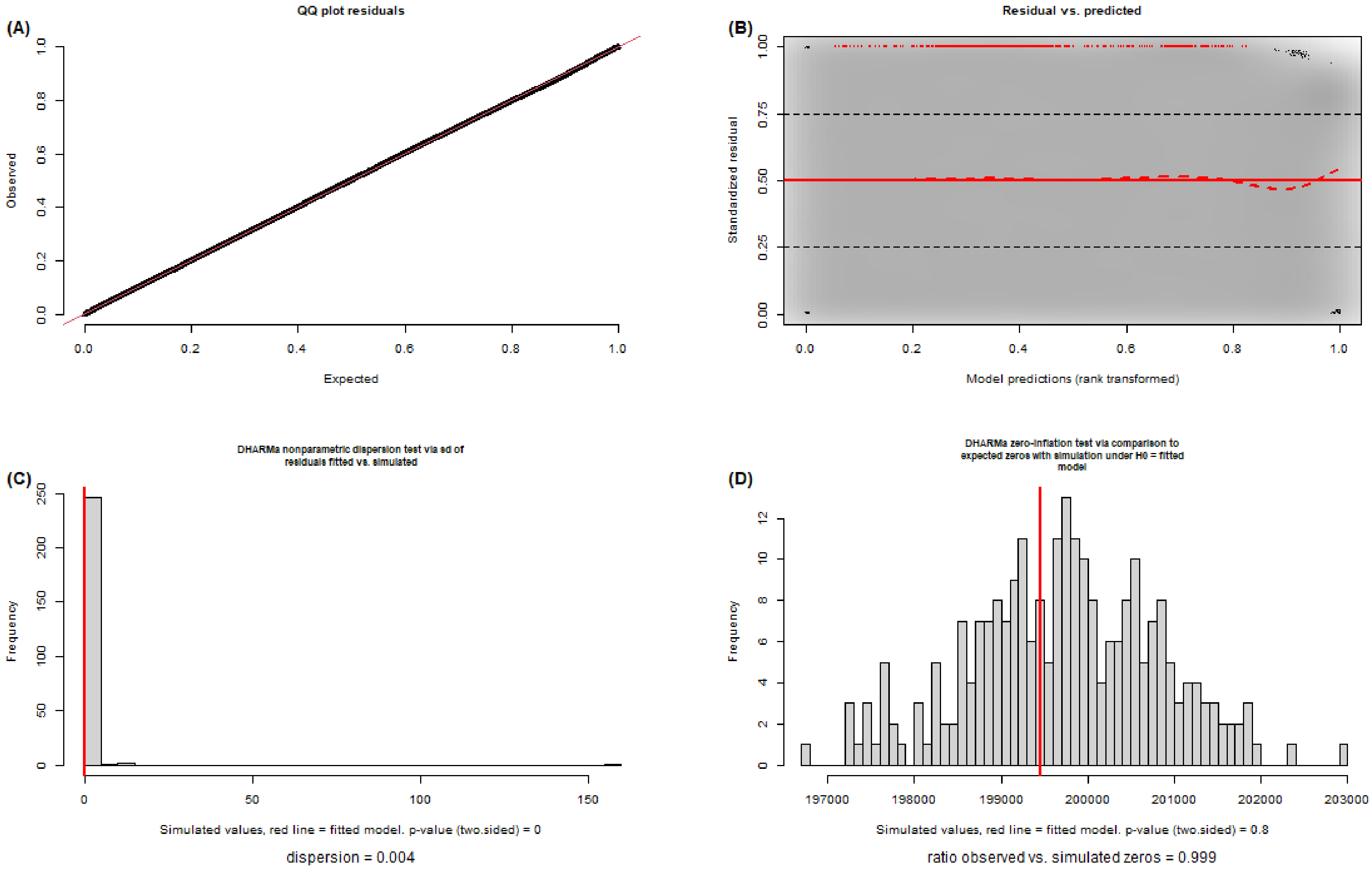


Figure 41. Diagnostics for the Swadesh-lists models (first corpus).

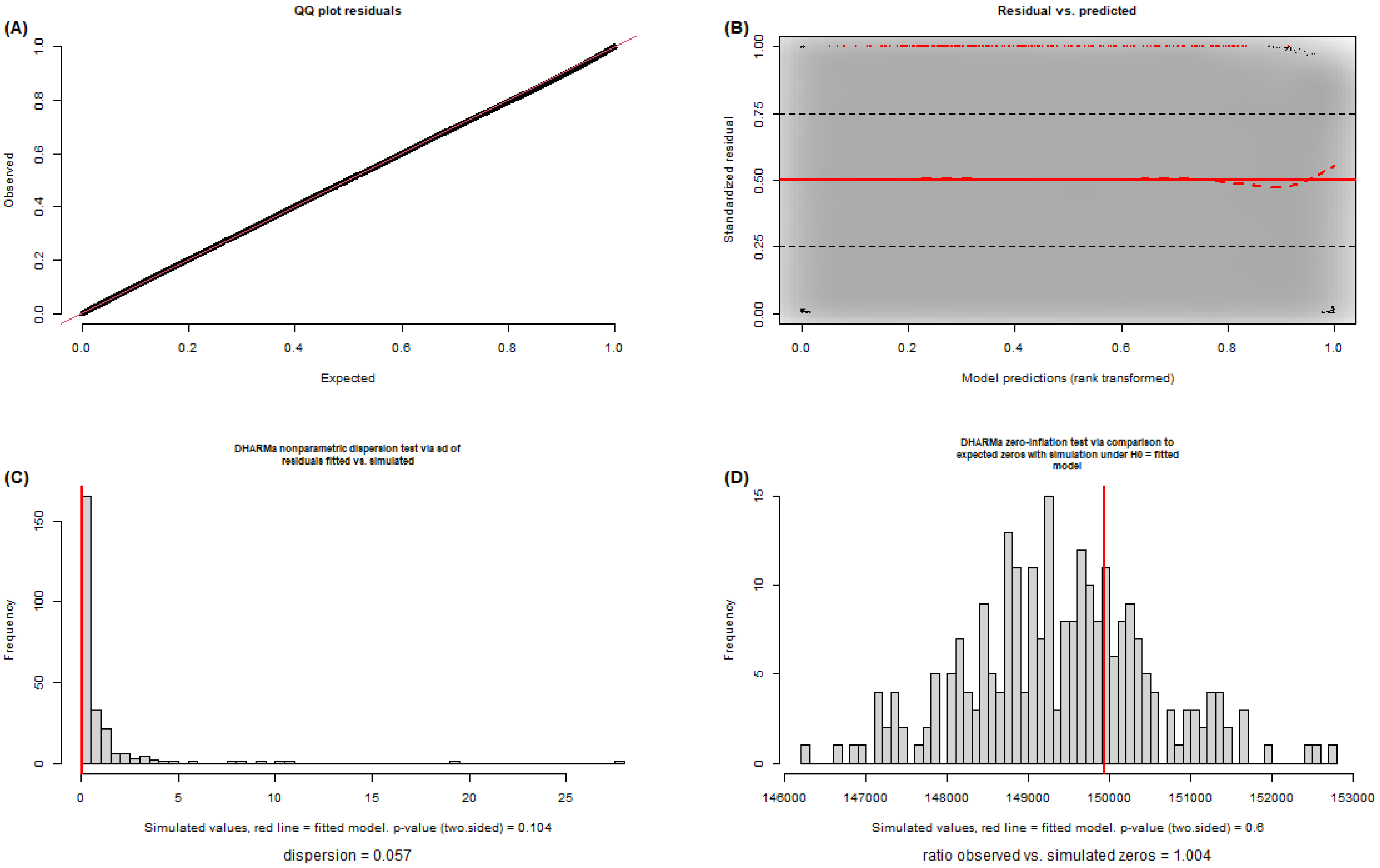


Figure 42. Diagnostics for the Swadesh-lists models (second corpus).

In the case of the parallel-dictionaries models, we were unable to run the full diagnostics on the full models, since the large size of the models necessitated memory allocation for the diagnostics that exceeded our available computational resources. To address this, we built new models using sub-samples from the original samples (separately for each corpus), containing 2,500,000 randomly selected observations each, and used these for the diagnostics.[131]

The results of these models, which are shown in Table 55, are similar to those of the main models, which supports their use for diagnostic purposes, though the model for the first corpus is slightly less well-specified than for the associated main model.[132] The results of the associated diagnostic checks, which appear in Figures 43 and 44, are similar to those of the Swadseh-based models, and suggest that the model are fairly well-specified, though they also have some underdispersion.

[131] The size of 2,500,000 observations was chosen since with a 3,000,000-observations sub-sample we still hit the memory allocation limit for the dispersion and zero-inflation tests.

[132] There are two key differences between the subsample-based model for the first corpus and the associated main model. First, this (subsample-based) model had a "singular convergence" warning, likely due to the random intercept for L1 and the associated random slope of distance for L1, though the associated effect sizes were very similar to those in the main models (i.e., functionally 0). Second, the frequency predictor in the subsample model is underestimated, as it has a smaller IRR (and SE) than in the main models, though the frequency predictor is still substantial. It is important to keep these differences in mind when it comes to the diagnostics, but they are nevertheless minor enough that this model is reasonable to use for diagnostic purposes, especially given that it is slightly less well-specified than the main model, which makes using it more conservative. In addition, note that, as expected, the differences between the subsample-based model and the main model generally become smaller as the size of the sub-sample increases, and the residual plots also become even closer to what is expected in a well-specified model. For example, when the sub-sample is increased to 3,000,000 observations, though there is still a "singular convergence" warning, the IRR and SE of frequency both become more similar to those of the associated main model (specifically, the IRR becomes 13.23 and the SE becomes 0.80), and the residual plot become even closer to what is expected for a well-specified model (i.e., the slight uptick at the right side of the plot flattens).

Table 55. Results of the mixed-effects models, for the parallel-based samples, using the 2,000,000-observation subsamples that were selected for diagnostics. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*
(Intercept)	-13.83	0.06	0.00	<0.01	-224.94	<.001	-12.52	0.05	0.00	<0.01	-236.96	<.001
Distance	0.01	<0.01	1.01	<0.01	2.54	.011	0.00	0.01	1.00	0.01	0.16	.871
Proficiency	0.02	0.01	1.02	0.01	2.16	.031	0.04	0.01	1.04	0.01	4.31	<.001
Frequency	2.21	0.06	9.15	0.57	35.60	<.001	2.91	0.05	18.34	0.96	55.74	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	0.61	.541	0.00	<0.01	1.00	<0.01	1.58	.115
Random effects												
Learner_τ_{00}	0.03						0.05					
Task_τ_{00}	0.03						0.08					
Word_τ_{00}	0.27						0.58					
Task:Word_τ_{00}	2.47						1.52					
L1_τ_{00}	0.00						0.00					
L1.Distance_τ_{11}	0.00						0.01					
L1_ρ_{01}	1.00						0.87					

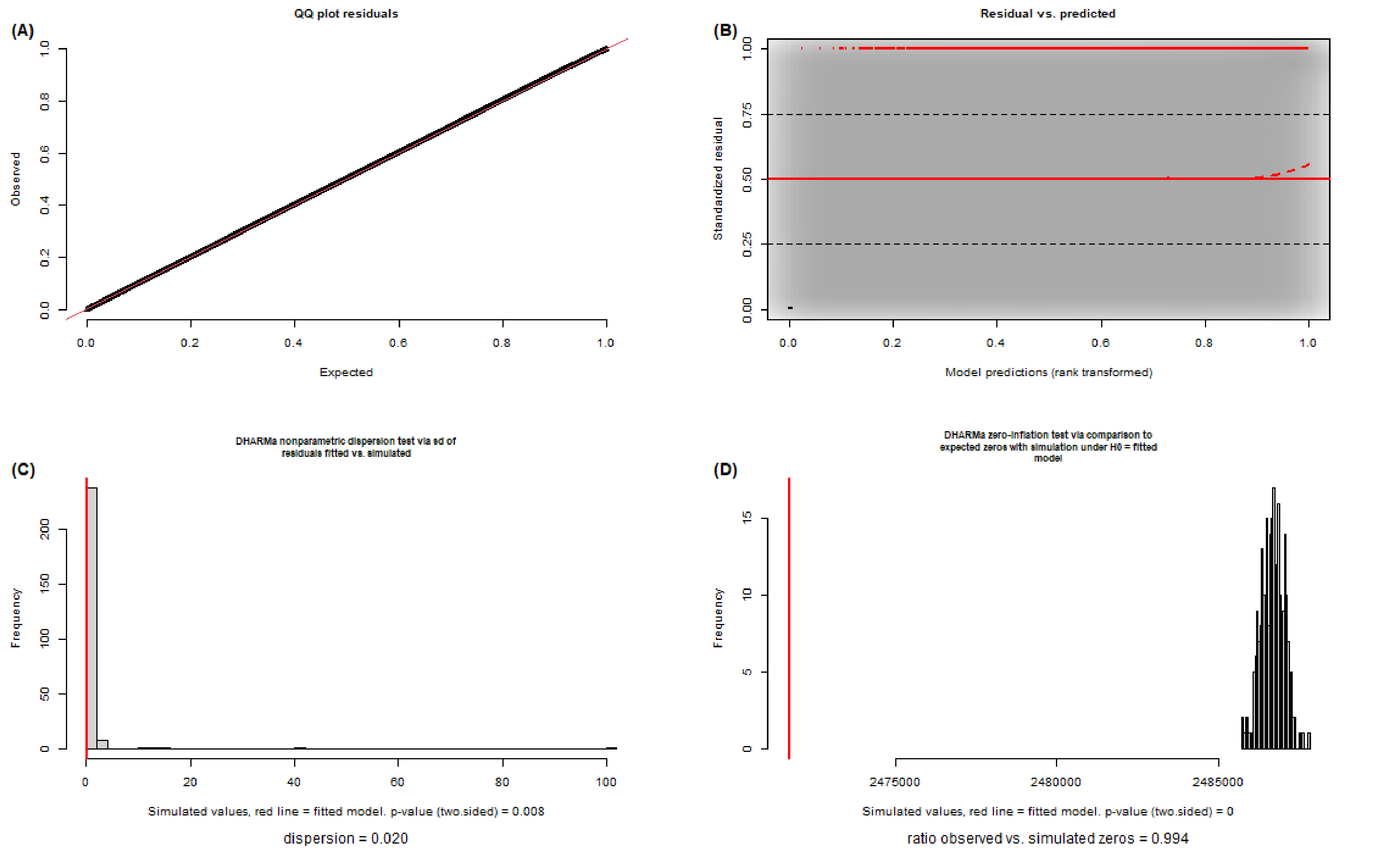


Figure 43. Diagnostics for the parallel-dictionaries models (first corpus). In the case of the zero-inflation test, note that the ratio between observed and simulated zeros is very close to 1, so this is not an issue for this model.

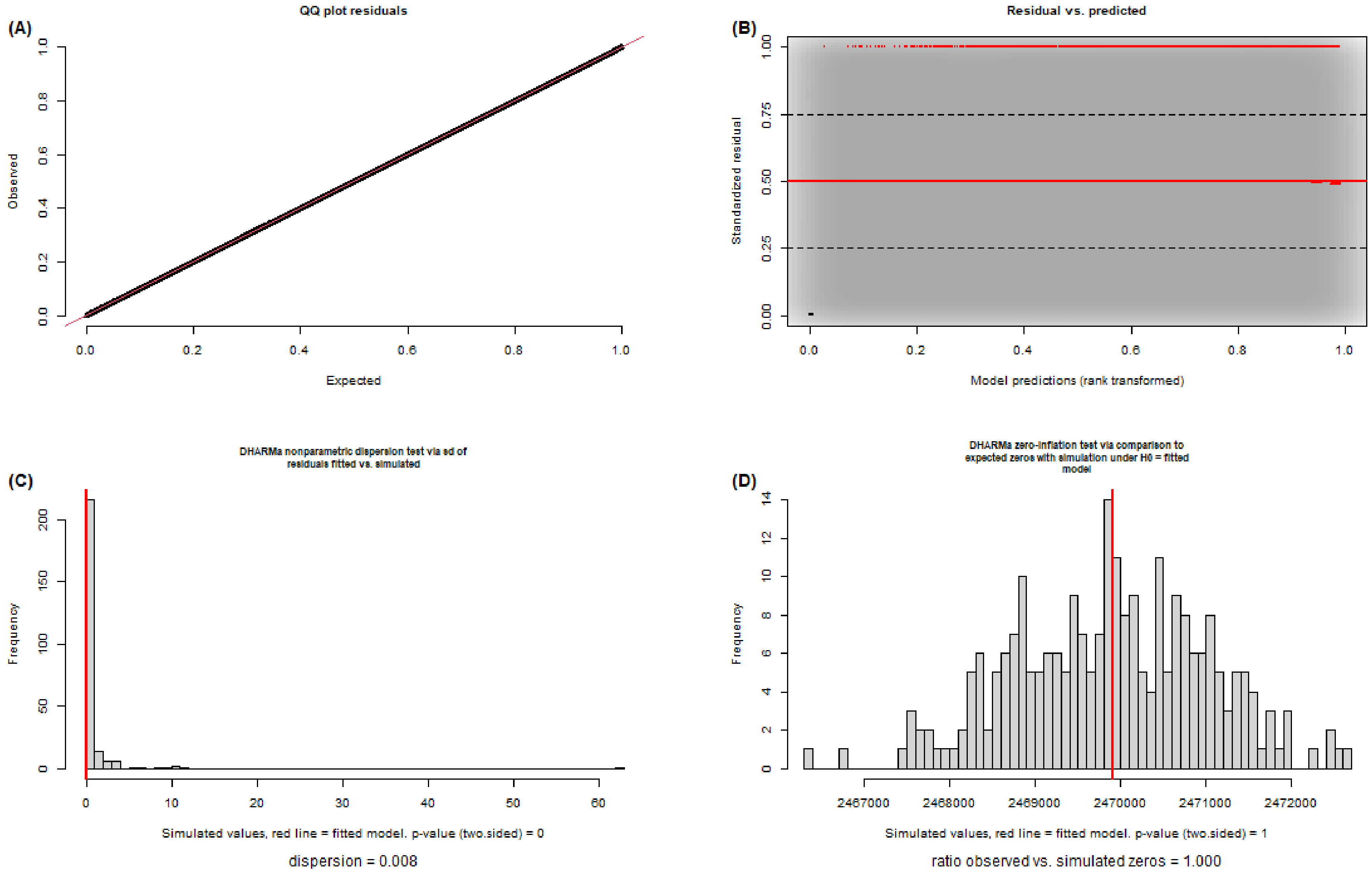


Figure 44. Diagnostics for the parallel-dictionaries models (second corpus).

Overall, the diagnostics for the models suggests that the models are fairly well-specified, though they have some underdispersion, which can lead to overestimated SEs, and consequently to overestimated p-values (Brooks et al., 2017, 2019; Dean & Lundy, 2016; Forthmann & Doebler, 2021; Harris et al., 2012; Hartig, 2021a; Sellers & Morris, 2017). However, this underdispersion does *not* invalidate the present findings, given the robust effect sizes that were found across all samples (IRRs very close to 1, with SEs ≤ 0.01), since even if these SEs are overestimated, the key patterns of results are still the same, in terms of the lack of effect of distance and of its interaction with proficiency. Essentially, even if these SEs should be smaller than they are, this would only reinforce our certainty regarding the estimated IRRs, and show that they are functionally equivalent to 1, which corresponds to a coefficient estimate of 0 and means that there is no effect. This is further supported by the supplementary models in the next sub-section, which replicate our findings while accounting for underdispersion. In sum, these diagnostics suggest that these models are fairly well-specified, and that they allow us to reliably answer our key research questions.

7.5.6.1.4 Supplementary models (generalized Poisson)

To account for any underdispersion in the main models, we built supplementary *generalized Poisson* models, which can handle both underdispersion and overdispersion (Brooks et al., 2019; Harris et al., 2012; Sellers & Morris, 2017; F. Zhu, 2012).[133] As shown below, these models suffered from various convergence issues, so they are not a viable option to use as the main models, and we do not compare them directly to the main models here in terms of performance (e.g., based on AIC/BIC). Nevertheless, these models had very similar results as the main models, which provides support for the key findings.

Specifically, Table 56 contains these models for the Swadesh-based samples. Both models had results that are extremely similar to the main models, particularly in the case of the

[133] In addition, we also attempted to build *Conway-Maxwell-Poisson* models, which can also handle both underdispersion and overdispersion (Brooks et al., 2017, 2019; Forthmann & Doebler, 2021; Lynch et al., 2014; Sellers & Morris, 2017). The reason for this attempt was that these models might be less prone to convergence problems, though they are also much more computationally intensive (Brooks et al., 2019). Unfortunately, they also had convergence warnings for the Swadesh-based model in the second corpus, similarly to the generalized Poisson models, so they were not helpful in this regard, and furthermore, due to their high computational costs, we were unable to get them to converge for the parallel-based samples. Nevertheless, this is not crucial, as the results for these models in the case of the Swadesh-based samples where they did converge were very close to those of the generalized-Poisson models, and functionally equivalent when it comes to the key variables under consideration (i.e., an IRR of 0.99–1 and an SE ≤.01 for *distance* and the *distance:proficiency* interaction).

key variables that the study focuses on (*distance* and the *distance:proficiency* interaction). The sample for the first corpus converged with a “NA/NaN function evaluation warning”.[134]

[134] See the glmmTMB documentation for a description and discussion of all the convergence warnings and errors mentioned here: http://web.archive.org/web/20210516105444/https://cran.r-project.org/web/packages/glmmTMB/vignettes/troubleshooting.html

Table 56. Results of the generalized Poisson models, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus						Second corpus					
Predictor	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*	*B*	*SE*$_B$	*IRR*	*SE*$_{IRR}$	*z*	*p*
(Intercept)	-10.28	0.12	0.00	<0.01	-86.73	<.001	-9.73	0.13	0.00	<0.01	-77.08	<.001
Distance	-0.01	0.01	0.99	0.01	-1.39	.165	-0.01	0.01	0.99	0.01	-0.59	.552
Proficiency	-0.05	0.02	0.95	0.02	-2.40	.016	-0.01	0.02	0.99	0.02	-0.67	.504
Frequency	3.29	0.14	26.78	3.78	23.32	<.001	3.09	0.16	21.95	3.52	19.29	<.001
Dist:Prof	0.00	<0.01	1.00	<0.01	0.54	.587	0.00	<0.01	1.00	<0.01	-1.04	.296
Random effects												
Learner_τ_{00}	0.00						0.17					
Task_τ_{00}	0.37						0.30					
Word_τ_{00}	0.37						0.47					
Task:Word_τ_{00}	1.81						1.29					
L1_τ_{00}	0.01						0.03					
L1.Distance_τ_{11}	0.01						0.02					
L1_ρ_{01}	0.76						-0.04					

Table 57 contains the generalized Poisson models for the parallel-based samples. There were more convergence issues here, as the first corpus did not converge at all (It had a “gradient function must return a numeric vector of length 13” error, as well as a “NA/NaN function evaluation” warning), and the second corpus converged with two warnings (“singular convergence” and a “non-positive-definite Hessian matrix”).[135] Nevertheless, the findings of the model that did converge, albeit with warnings, are very similar to those of the associated main model.

[135] As noted previously, see the glmmTMB documentation for a description and discussion of all the convergence warnings and errors mentioned here: http://web.archive.org/web/20210516105444/https://cran.r-project.org/web/packages/glmmTMB/vignettes/troubleshooting.html

Table 57. Results of the generalized Poisson models, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *distance* is the phonological LDN between each L2 word and its most lexically similar L1 counterpart (originally 0–1, scaled to 0–10), *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} and τ_{11} respectively represent the SD of the associated random intercepts and slopes, and ρ_{01} represents the correlation between random intercepts and associated random slopes (here, *distance* for *L1*).

	First corpus [a]						Second corpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)							-12.49	0.04	0.00	<0.01	-289.28	<.001
Distance							0.01	0.01	1.01	0.01	1.33	.184
Proficiency							0.04	0.01	1.04	0.01	7.09	<.001
Frequency							2.95	0.04	19.07	0.84	67.23	<.001
Dist:Prof							0.00	<0.01	1.00	<0.01	1.11	.265
Random effects												
Learner_τ_{00}							0.00					
Task_τ_{00}							0.11					
Word_τ_{00}							0.67					
Task:Word_τ_{00}							1.43					
L1_τ_{00}							0.00					
L1.Distance_τ_{11}							0.01					
L1_ρ_{01}							1.00					

[a] There are no results for the model in the first corpus since it did not converge, but the table is kept in the same format as for the other models to facilitate comparisons.

In summary, we attempted to build models that use variants of the Poisson distribution that can handle both underdispersion and overdispersion (namely, generalized Poisson models). The resulting models had a number of convergence issues, errors, and warnings, which supports the use of the regular Poisson models as the main models in the study. Nevertheless, the findings in the models that did converge, including those that converged with no warnings (i.e., the Swadesh-based models in the first corpus) mirror the findings of the main models, especially with regard to the key variables in the study (the *distance* predictor and the *distance:proficiency* interaction). This was expected, since the main issue with underdispersion are overestimated SEs (Brooks et al., 2017, 2019; Dean & Lundy, 2016; Forthmann & Doebler, 2021; Harris et al., 2012; Hartig, 2021a; Sellers & Morris, 2017), and this is not a problem here, given the very small SEs that were found across all samples. As such, these models provide support for the findings of the main models, and suggest that any potential underdispersion in the data does not substantially change our key findings.

7.5.6.2 Collinearity

In addition to residual plots, we checked for potential collinearity using the *performance* package in R (Lüdecke et al., 2021).[136] The results of this appear in Figure 45, which contains the *variance inflation factor* (VIF) for the predictors in each model. In all cases, the VIF was minimal (i.e., equal to or very close to 1), which indicates the collinearity was not an issue for the present analyses, especially given the large sample sizes (Morrissey & Ruxton, 2018; R. M. O’Brien, 2007; Winter, 2019).

[136] The VIF values were calculated using the *performance* package, and the results were plotted using the base R *barplot* function.

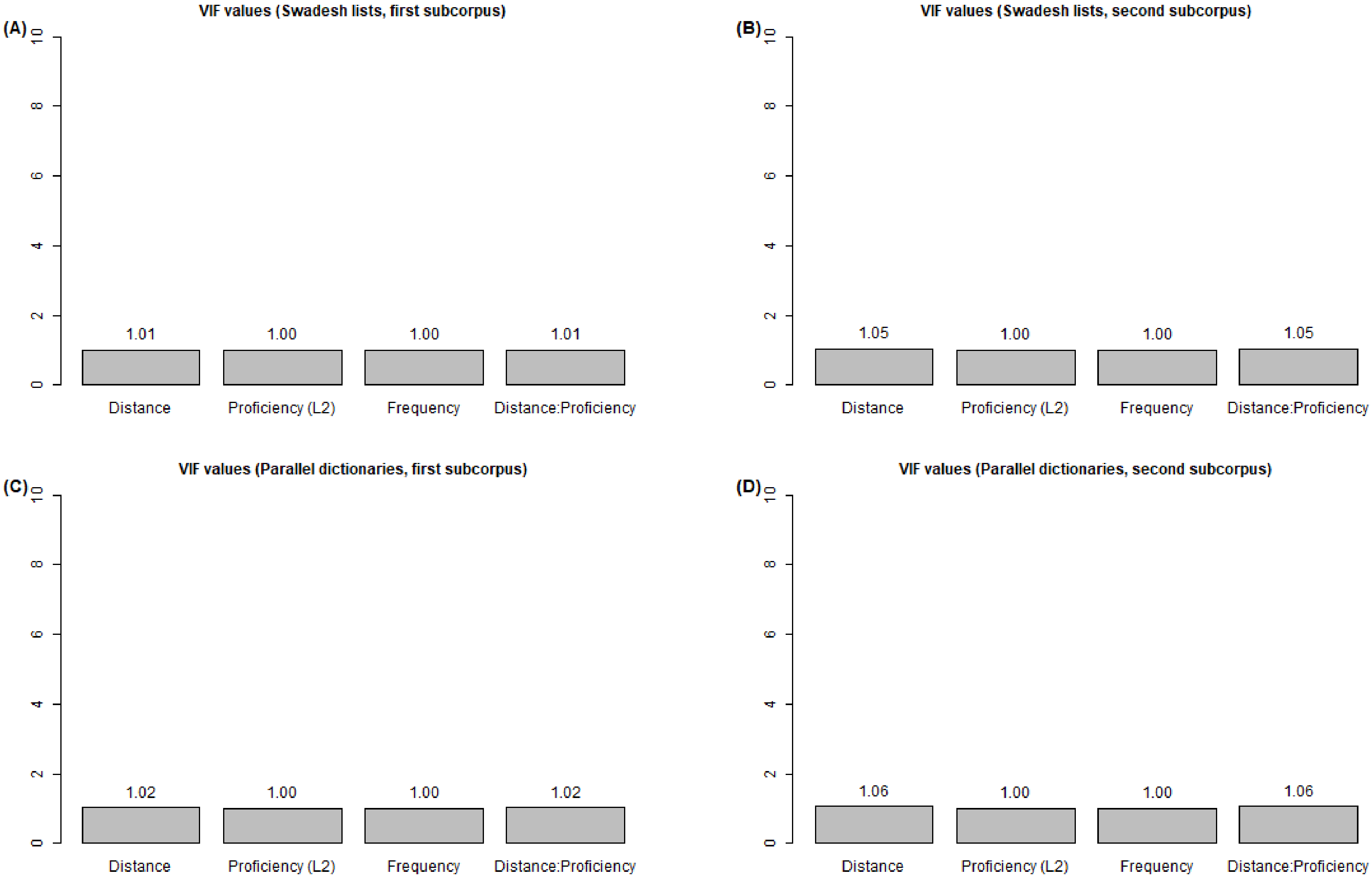


Figure 45. Plots showing the *variance inflation factor* (VIF) for the predictors in each sample, to check for collinearity.

7.5.7 Baseline models (without distance)

The baseline models were models that did not include lexical distance at all (MODELS$_{\text{baseline}}$). We compared these models to the main models that were used in the study (MODELS$_{\text{main}}$), where lexical distance was included as a predictor, as part of an interaction with L2 proficiency, and as random slopes of L1. In addition, to better understand how the removal of lexical distance from the models influences them,[137] we also compared the baseline and main models with models that contained distance as a predictor/interaction but without random slopes (MODELS$_{\text{no\_slope}}$), and with models that had distance only as a predictor, with no random slopes or interaction (MODELS$_{\text{only\_predictor}}$).

Specifically, we compared the models' AIC and BIC, and the results of this are shown in Table 58. Both measures were used, as suggested in Kuha, (2004). The AIC and BIC of each model were extracted directly from each model object in R using the *summary* function. All comparisons were between models that used the same set of data (i.e., between models that use the same learner sample and lexical-distance dataset), as required when using these measures (Fabozzi et al., 2014; Kuha, 2004).

[137] For example, this could show if removing lexical distance from the interaction improves the models, but removing it as a predictor worsens them.

Table 58. Comparisons of AIC and BIC across models.

Corpus	Distance data	Model	AIC	Δ AIC	BIC	Δ BIC
First	Swadesh	baseline	73703.03	0.58	73785.16	-
First	Swadesh	only predictor	73702.45	-	73794.85	9.69
First	Swadesh	no slope	73704.03	1.59	73806.70	21.54
First	Swadesh	main	73705.32	2.87	73828.52	43.36
Second	Swadesh	baseline	61475.10	5.66	61554.95	-
Second	Swadesh	only predictor	61474.56	5.11	61564.39	9.44
Second	Swadesh	no slope	61474.99	5.55	61574.80	19.86
Second	Swadesh	main	61469.44	-	61589.22	34.27
First	parallel	baseline	401614.39	-	401722.16	-
First	parallel	only predictor	401663.18	48.80	401784.42	62.27
First	parallel	no slope	401662.38	48.00	401797.09	74.94
First	parallel	main	401655.51	41.12	401817.16	95.01
Second	parallel	baseline	346322.57	5.27	346428.02	-
Second	parallel	only predictor	346322.72	5.41	346441.34	13.33
Second	parallel	no slope	346323.49	6.18	346455.29	27.28
Second	parallel	main	346317.31	-	346475.47	47.46

Note. ΔAIC is calculated by subtracting the AIC of a given model from the AIC of the model with the minimal AIC for that combination of corpus (i.e., first/second) and lexical-distance dataset (i.e., Swadesh/parallel), since comparisons can only be made between models that are based on the same data (Fabozzi et al., 2014; Kuha, 2004). Accordingly, no ΔAIC is listed for the model with the minimal AIC for a certain combination (e.g., Swadesh lists in the first corpus). The same is the case for ΔBIC.

Interpretations of the differences in AIC/BIC are based on Fabozzi et al. (2014). In terms of BIC, there was very strong support for the simplest (baseline) model in all 4 cases, as it had the minimal BIC, with ΔBIC either slightly below 10 or far above it. In terms of AIC, the picture was less clear. Specifically, in the case of the parallel dictionaries in the first corpus, the baseline model was strongly supported (ΔAIC > 40). However, in the case the first corpus in the Swadesh lists, there was only weak support for the baseline and predictor-only models over the main model (ΔAIC ~2–3), and in the case of the second corpus (both Swadesh and parallel), there was moderate support (ΔAIC ~5–6) for the main models over the other models (though the main models were ranked the worst in all cases based on BIC). This difference between AIC/BIC can be attributed to the greater penalty that BIC imposes for the number of parameters in the model (Fabozzi et al., 2014). When the patterns of the two measures are considered, together with the estimates for the associated predictors, it appears that the AIC comparisons are sometimes recommending the use of an overfitted model here.

Overall, the comparison between the models did not consistently support the inclusion of linguistic distance as a predictor based on AIC, and consistently supported its exclusion based on BIC. It is, therefore, reasonable to conclude that the effect of distance is at best unclear in our dataset. This is strongly supported by the findings for the main models that are shown in the paper, where the distance predictor and the interaction had IRRs very close to 1 (corresponding to a coefficient estimate of 0) and very small SEs, and where the SDs of the random slopes of distance were also very close to 1 (i.e., to a coefficient estimate of 0).

The results for the baseline models are shown in Tables 59 and 60.

Table 59. Results of the baseline mixed-effects models, for the Swadesh-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the standard deviation (SD) of the associated random intercepts.

	First corpus						Second corpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-10.32	0.16	0.00	<0.01	-65.30	<.001	-9.85	0.14	0.00	<0.01	-68.62	<.001
Proficiency	-0.04	0.02	0.96	0.02	-2.11	.035	0.00	0.02	1.00	0.02	-0.25	.806
Frequency	3.29	0.21	26.89	5.65	15.68	<.001	3.15	0.19	23.31	4.44	16.52	<.001
Random effects												
Learner_τ_{00}	0.07						0.23					
Task_τ_{00}	0.40						0.33					
Word_τ_{00}	0.38						0.46					
Task:Word_τ_{00}	1.84						1.36					
L1_τ_{00}	0.02						0.03					

Table 60. Results of the baseline mixed-effects models, for the parallel-based samples. The response variable was the rate of use of the target L2 English words (i.e., their count offset by the total number of words in each text). Under *fixed effects*, *proficiency* is the EFCAMDAT L2 proficiency level at which the text was written (1–12, corresponding to CEFR A1–B2), and *frequency* is the baseline Zipf frequency of the target word in English (~1–7.5). Under random effects, τ_{00} represents the standard deviation (SD) of the associated random intercepts.

	First corpus						Second corpus					
Predictor	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*	*B*	SE_B	*IRR*	SE_{IRR}	*z*	*p*
(Intercept)	-12.86	0.06	0.00	<0.01	-207.24	<.001	-12.59	0.05	0.00	<0.01	-242.46	<.001
Proficiency	0.11	0.01	1.11	0.01	9.02	<.001	0.04	0.01	1.04	0.01	4.36	<.001
Frequency	2.90	0.06	18.14	1.05	49.90	<.001	2.97	0.05	19.57	0.99	58.57	<.001
Random effects												
Learner_τ_{00}	0.03						0.05					
Task_τ_{00}	0.03						0.12					
Word_τ_{00}	0.46						0.65					
Task:Word_τ_{00}	2.30						1.50					
L1_τ_{00}	0.01						0.01					

7.5.8 Analysis of synonym sets

Rabinovich et al. (2018) use a different approach than us to analyze the cognate facilitation effect, with the key difference being that they examined how cognancy within a synonym set (*synset*) influences word choice, whereas we examined how similarity between words and synonym sets influences word choice. While conducting a similar style of analysis as them is beyond the scope of this study, below we present a brief analysis of our sample in light of the approach used by Rabinovich et al..

Based on our own research and on the conditions in which Rabinovich et al. (2018) found evidence of the cognate facilitation effect, we can identify the following criteria as conditions under which we would expect to find this effect within synonym sets:

- **There must be a communicative need or reasonable opportunity to convey the relevant meaning.** They characterize their sample as involving "spontaneous productions", so in their case it is likely that learners had more opportunities for choosing which meanings to convey than in more constrained task-based settings.
- **The relevant meaning must be able to be conveyed using a synset.** This is because the cognate facilitation effect, as found by them, is based on the contrast in usage between synonyms within a synset.
- **The synonyms must be easily interchangeable.** This is because otherwise, the effects of cognancy may be obscured by other factors that play a role in the choice of specific synonyms out of the synset, and especially frequency effects. In their study, they operationalized this concept by avoiding synsets that were dominated by a single synonym (i.e., where a single synonym accounted for 90% or more of the usage of that synset in their dataset). This means, for example, that a synset such as {*kiss*, *buss*, *osculation*} was excluded, whereas a synset such as {*divide*, *split*} was retained.[138]
- **There must be a mix of cognates and non-cognates in the synset.** Specifically, there must be at least one cognate for the cognate facilitation effect to occur, but there must also

[138] While this is a reasonable operational definition from a practical perspective, especially when working with large-scale datasets, it is important to note that there are various issues with it. For example, some synonyms might not be easily interchangeable due to connotations that they carry, even if they have a similar rate of usage. In addition, the reliance on a strict 90% threshold can lead to issues, such as in a case where a single synonym accounts for 85% of the uses in a corpus, meaning that it is still fairly dominant over the others. Similarly, there can be a difference between a synset with two synonyms that each account for 50% of uses, and a synset with 3 synonyms that has a usage distribution of 50%-49%-1% or 50%-25%-25%. Finally, if a certain L2 word a cognate in many languages, it might become a highly dominant synonym, and therefore be omitted from the sample even though it displays a strong cognate facilitation effect.

be at least one non-cognate against which the cognate stands out.[139] Note that this criterion is L1-dependent, since cognancy of an L2 word is defined based on its relation to an L1 word.

We briefly analyzed our samples to determine to what degree these conditions occur there.

In the Swadesh lists, none of the English words were listed as being a part of a synset. In the parallel dictionaries, out of 1,103 English words that were included in our analyses, 751 (68.09%) were listed as having no synonyms, and 352 (31.91%) were part of a synset. Of those with a synonym, 21 (5.97%) of the entries that originally had two synonyms in the dataset appeared by themselves in the final dataset, due to removal of the other synonym in during the data preparation.[140] Of the 331 entries that were a part of a synset in the present dataset, 304 (91.84%) were part of a synonym pair (i.e., a synset with 2 synonyms), and 27 (8.16%) were a part of a synonym triplet (i.e., a synset with 3 synonyms). As such, there were a total of 161 synsets in our parallel-dictionaries dataset.

When considering how many of these were easily interchangeable, we based our criterion on a similar one as Rabinovich et al., and define an easily interchangeable synset as one where the difference in Zipf frequency between the synonyms is no greater than 1 (i.e., where no synonym is 10 times or more common than the others, since Zipf frequency is on a logarithmic scale). 110 (68.32%) of the synsets (corresponding to 223 entries) fulfilled this criterion, with Zipf frequency differences ranging all the way from 0.00–0.99.

Next, there was the question of which of these synsets contain a difference in lexical similarity that could be characterized as corresponding to cognancy/non-cognancy, since we use a continuous measure of lexical similarity, rather than something that clearly delineates whether a pair of words are cognates or not. As a rough measure, we categorized synset as fulfilling this criterion if at least one of the synonyms had an LDN ≤ .60 and at least one had an LDN ≥ .80.[141] Unlike the previous criteria, which were L1-independent, this was L1-dependent, so there were 550 relevant synset combinations (110 synsets for each of the 5 L1s

[139] However, it may also be the case that there can be a facilitative effect of lexical similarity even if there are no cognates in a synonym set, as long as some of the synonyms are substantially more similar to the L1 counterpart than the other synonyms are.

[140] The remaining entries with synonyms did not have any of their synonyms removed during the data-preparation stage.

[141] As with the 90% frequency cutoff proposed by Rabinovich et al., the exact cut-off that was chosen is somewhat arbitrary, and any single cutoff that is used will likely involve a tradeoff between false positives and false negatives. The specific values that were chosen here are based on a manual examination of the data, and while arguments could be made for other values or criteria, it does not appear that this would substantially change the findings of this analysis.

in the parallel dictionaries). Of these, 93 (16.91%), which contain 189 synonyms, fulfilled the cognancy criterion.

Finally, there was the question of whether there was a communicative need for the underlying meanings represented by these synsets. This was determined based on whether at least one of the synonyms in the relevant synsets appeared at least once in a text:

- In the first corpus, there were 179,439 rows which represent a combination of one of the above synonyms with a text (while taking learners' L1 into account). Of these, 710 (0.4%) rows had a count > 0 for the target word, meaning that it was used at least once.[142] These represented the use of 63 synsets (67.74% of the original synsets).
- In the second corpus, there were 134,190 rows which represent a combination of one of the above synonyms with a text. Of these, 709 (0.53%) rows had a count > 0 for the target word. These represented the use of 64 synsets (68.82% of the original synsets).

Overall, this suggests that, in the present samples, there was a substantial number of cases where words were used from a synset that fulfills the necessary criteria for the cognate facilitation effect (interchangeability, a combination of cognancy/non-cognancy, and a communicative need for the underlying meaning).

This aspect of the data should be interpreted with caution, since there are various issues with how these criteria are operationalized and with how synonyms are listed in the datasets in the first place. For example, there are cases where synonyms that fit these criteria are not really interchangeable, as in the case of {*vein*/*artery*}, or are only interchangeable in some situations, as in the case of {*marriage/wedding*}. Nevertheless, even taking such issues into account, it seems that at least some of the entries in the present analyses include cognates as part of a fairly interchangeable synset (e.g., {*woods*/*forest*}, {*stone*/*rock*}, {*carriage*/*wagon*/*cart*}), so it may be possible to use this sample in analyses that are similar to those of Rabinovich et al..

7.5.9 Software used in the analyses

All analyses were performed in R (R Core Team, 2021).[143] All tests of statistical significance throughout the study were two-tailed. To list the specific packages that were loaded throughout

[142] See §7.5.17.5.3 for summary statistics of the usage of the target words. Briefly, we can say that this apparently low rate is expected, since many words are used rarely, when there is a need for them in a task.

[143] However, the lexical-distance data was generated in Python. Specifically, the following Python libraries were used for basic data wrangling and calculations: *SciPy* (Virtanen et al., 2019), *pandas* (McKinney, 2010), and *numpy* (Oliphant, 2006; Walt et al., 2011). The ASJP's phonetic script (outlined in Brown et al., 2008) was

the analyses, we used the *sessionInfo* function from the *report* library (Makowski & Lüdecke, 2019). This generates an automated output based on the citation information associated with the metadata of each package, which may be incomplete or formatted differently than APA style. We present this bibliography here as-is, to preserve the original output, and we therefore also separate it from the main the main bibliography for this document.

---*Start of report(sessionInfo()) output below*---

Analyses were conducted using the R Statistical language (version 4.0.4; R Core Team, 2021) on Windows 10 x64 (build 19042), using the packages broom.mixed (version 0.2.6; Ben Bolker and David Robinson, 2020), DHARMa (version 0.4.1; Florian Hartig, 2021), ggplot2 (version 3.3.3; Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.), stringr (version 1.4.0; Hadley Wickham, 2019), forcats (version 0.5.1; Hadley Wickham, 2021), tidyr (version 1.1.3; Hadley Wickham, 2021), readxl (version 1.3.1; Hadley Wickham and Jennifer Bryan, 2019), readr (version 1.4.0; Hadley Wickham and Jim Hester, 2020), dplyr (version 1.0.5; Hadley Wickham et al., 2021), tibble (version 3.1.0; Kirill Müller and Hadley Wickham, 2021), purrr (version 0.3.4; Lionel Henry and Hadley Wickham, 2020), sjPlot (version 2.8.7; Lüdecke D, 2021), performance (version 0.7.0; Lüdecke et al., 2020), glmmTMB (version 1.0.2.1; Mollie Brooks et al., 2017), openxlsx (version 4.2.3; Philipp Schauberger and Alexander Walker, 2020) and tidyverse (version 1.3.0; Wickham et al., 2019).

References

- Ben Bolker and David Robinson (2020). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.6. https://CRAN.R-project.org/package=broom.mixed

- Florian Hartig (2021). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.4.1. https://CRAN.R-project.org/package=DHARMa

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

converted to IPA using the *asjp* library (Sofroniev, 2018). Distances were calculated using the *PanPhon* library (Mortensen et al., 2016).

- Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

- Hadley Wickham (2021). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. https://CRAN.R-project.org/package=forcats

- Hadley Wickham (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

- Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

- Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr

- Kirill Müller and Hadley Wickham (2021). tibble: Simple Data Frames. R package version 3.1.0. https://CRAN.R-project.org/package=tibble

- Lionel Henry and Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr

- Lüdecke D (2021). _sjPlot: Data Visualization for Statisticsin Social Science_. R package version 2.8.7, <URL:https://CRAN.R-project.org/package=sjPlot>.

- Lüdecke, Makowski, Waggoner & Patil (2020). Assessment of Regression Models Performance. CRAN. Available from https://easystats.github.io/performance/

- Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler and Benjamin M. Bolker (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. The R Journal, 9(2), 378-400.

- Philipp Schauberger and Alexander Walker (2020). openxlsx: Read, Write and Edit xlsx Files. R package version 4.2.3. https://CRAN.R-project.org/package=openxlsx

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

---*End of report(sessionInfo()) output above*---