The Potential Influence of Crosslinguistic Lexical Similarity on Lexical Diversity in L2 English Writing

Article accepted in *Corpora*

Itamar Shatz[1*], Theodora Alexopoulou[1], Akira Murakami[2]

[1] University of Cambridge; [2] University of Birmingham

* Corresponding author, is442@cam.ac.uk

Abstract

We examined the potential influence of L1-L2 lexical similarity on L2 lexical diversity, to determine whether the robust facilitative effect of lexical similarity that is found in processing and broad learning outcomes extends to this measure of L2 production. Our sample included two matching learner sub-corpora, containing 8,500 and 6,390 English texts, written in response to 95 and 71 writing tasks, by speakers of 9 typologically diverse L1s, in the A1–B2 CEFR range of L2 English proficiency. We found that lexical similarity did not influence L2 lexical diversity, at any proficiency level. This finding suggests that the facilitative effect of similarity does not necessarily extend to L2 production, at least in the case of certain global measures, like lexical diversity, and certain task-based settings, like the present educational one, where lexical choices are driven primarily by the constrained communicative needs of the tasks. This is supported by the strong task effects that we found, which we quantified using mixed-effects models, while also shedding light on the use of lexical diversity as an indicator of L2 proficiency.

Keywords: lexical transfer, lexical diversity, lexical similarity, language distance, crosslinguistic influence, learner corpus, vocabulary, second language acquisition

Acknowledgements: This work is based on the PhD thesis of I.S. (Shatz, 2022). I.S. acknowledges financial support from Hughes Hall at the University of Cambridge and Cambridge Assessment English; T.A. acknowledges support from the Isaac Newton Trust at the University of Cambridge and EF Education First. The funders had no role in the decisions made in this study.

1 INTRODUCTION

People's native language (L1) can influence their knowledge and use of a second language (L2), a phenomenon known as *crosslinguistic influence* or *language transfer*. This can facilitate language acquisition/use, in which case it is sometimes called *positive transfer*, or hinder language acquisition/use, in which case it is sometimes called *negative transfer* or *interference* (Jarvis, 2017; Jarvis & Pavlenko, 2008; Ringbom, 2007). Transfer can occur in various linguistic domains, and here we focus on *lexical transfer*, which occurs when people's knowledge and use of words in one language influences their knowledge and use of words in another language (Jarvis et al., 2012; Jarvis & Pavlenko, 2008; Ringbom, 2007). A prominent example of such transfer is the *cognate facilitation effect*, a form of positive transfer that makes it easier for learners to process and learn an L2 word if it is psycholinguistically *cognate* with its L1 translation, meaning that it is similar to it phonologically and/or orthographically (De Wilde et al., 2020; Jarvis, 2009; Ringbom, 2007; Williams, 2015). This effect means, for instance, that it will generally be easier for English speakers to acquire the word for "lemon" in Italian ("limone"), than the word for "apple" ("mela").

The general explanation for this well-established phenomenon in processing and learning is that increased L1-L2 similarity in form facilitates the linking and mapping of L2 words to their L1 counterparts or to concepts that are shared between the languages, which facilitates the transfer of linguistic (e.g., syntactic, semantic, and morphological) information from the L1 to the L2 for the associated words (Jarvis & Pavlenko, 2008; Ringbom, 2007). In addition, there is general agreement that lexical transfer from learners' L1 plays a greater role at earlier stages of L2 acquisition (Jarvis, 2009; Williams, 2015). Overall, as Ringbom (2007, p. 28) states:

> 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.

However, while there is strong support for the cognate facilitation effect in the context of L2 processing, there is limited research on its role in L2 production. This creates uncertainty regarding the scope of this important effect, though researchers are beginning to investigate it more (Rabinovich et al., 2018).

There is evidence that increased lexical similarity between learners' L1 and their target L2—as measured based on the mean similarity between corresponding L1-L2 word pairs—can facilitate overall L2 acquisition, as measured using L2 proficiency scores (Schepens et al., 2013, 2020; van der Slik, 2010). However, this research relied on composite L2 proficiency scores, which includes multiple components such as grammar and coherence, so it is unclear how exactly the facilitative effect of crosslinguistic similarity influenced learners' L2. There is also evidence that learners' L1 can influence their choice of L2 words in language production, a phenomenon known as *word choice transfer* (Jarvis et al., 2012; Jarvis & Pavlenko, 2008; Kyle et al., 2015; Rabinovich et al., 2018). However, the limited research on this type of transfer does not make it clear whether it can be attributed to crosslinguistic similarity, particularly in relatively constrained task-based settings (e.g., prompt-based essays in language tests), as opposed to more spontaneous settings (e.g., social media). This is important, because task-based settings are common in education and assessment, and because word choice in such settings might be more strongly influenced by communicative needs and task effects, which could potentially interfere with the facilitative effect of crosslinguistic similarity.

Indeed, Crossley and McNamara (2011) suggest that this effect 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* (Granger et al., 2002), written in response to one of a 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*.[1] Lexical diversity—which we explain from a more technical perspective in §2.3—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 often used in language assessment and research as a key indicator of learners' L2 vocabulary and of their ability to use it, and it generally increases as learners' overall L2 proficiency increases, though the correlation between lexical diversity and L2 proficiency is imperfect

[1] The key lexical diversity measure that Crossley and McNamara (2011) used in their analyses is *Maas*, which they refer to as *M* (Maas, 1972).

(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, can elicit different average lexical diversity than describing the plot of a movie, due to factors like the difference in topic between these tasks (Alexopoulou et al., 2017; Johnson, 2017; Kessler et al., 2022; Torruella & Capsada, 2013; Yoon, 2017; Yu, 2010; Zenker & Kyle, 2021). The strong influence of tasks on lexical diversity supports the suggestion that task-based settings may override the influence of lexical similarity on lexical diversity. 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 address the key limitations that they mention, and follow their call to replicate and extend their research, to shed light on this important aspect of crosslinguistic influence during L2 acquisition. In addition, given the importance of lexical diversity in language assessment and research, and given that it should reflect a global cognate facilitation effect, if there is one, we focus on this measure in particular in our study. We 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. Furthermore, 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 this 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. This is beneficial, since past studies that examined these effects generally did not focus on them and/or used narrower samples, and is important given the practical applications of lexical diversity in language assessment and research.

Based on all this, we investigate the following in our study:

1. Does crosslinguistic lexical similarity influence L2 lexical diversity in a task-based educational setting, either in general or at low proficiency levels? This is the main research question, and the findings of Crossley and McNamara (2011) suggest that there might not be such crosslinguistic influence. However, it is important to investigate this, given the robust influence of such similarity in other settings and on other measures (e.g., De Wilde et al., 2020; Jarvis et al., 2012; Rabinovich et al., 2018; Schepens et al., 2020)., which suggests that similarity might play a role here too.

2. How well does L2 proficiency function as an indicator of lexical diversity? We expect lexical diversity to increase with proficiency, but this correlation is likely to be imperfect (Alexopoulou et al., 2017; Treffers-Daller et al., 2018), and we want to quantify its magnitude, as well as the variance in lexical diversity within proficiency levels.

3. How much does lexical diversity vary across different writing tasks? We expect strong task effects, but are uncertain regarding their exact magnitude. Note that (as explained in §2.1 and §2.4), we operationalize "task effects" in a sense that is rooted in our statistical approach to analysis, which utilizes mixed-effects models to enable us to conduct certain types of large-scale analyses. This operationalization is conventional when using such models (Winter, 2019), but differs from some other ways that "task" and "task effects" are defined in certain contexts (Alexopoulou et al., 2017).

Since lexical diversity is influenced by learners' L2 proficiency, it is necessary to control for this factor when comparing the writing of learners with different L1s. Past studies (e.g., Crossley & McNamara, 2011) generally did this by looking at learners who all come from roughly the same proficiency level. We use a similar but different approach, and control for this statistically, by including L2 proficiency as a predictor in our models. This allows us to include learners with a range of L2 proficiency levels in our analyses, and consequently to examine the potential interaction between L2 proficiency and the influence of similarity. Although the idea that learners at the same proficiency level can display different lexical diversity in their writing may be counter-intuitive. For example, it might be expected that L1 effects will only influence the rate of which learners' lexical diversity develops, and therefore the rate at which their L2 proficiency increases, rather than their lexical diversity at any given L2 proficiency levels.

However, past studies that used a similar approach as us and the same learner sample (the *EFCAMDAT*) found L1 effects in other measures, like the accuracy of grammatical morphemes (Murakami, 2016), as well as in many other linguistic phenomena, including relative clauses (Alexopoulou et al., 2015), clause-initial prepositional phrases (Jiang et al., 2014), capitalization (Shatz, 2019), and preference for certain phrases (Jiang et al., 2014). Furthermore, there is also evidence of similar L1 effects in other samples, as in the case of morpheme accuracy in the *Cambridge Learner Corpus* (Murakami & Alexopoulou, 2016). These measures are somewhat different from the one we focus on, for example because they focus on morphosyntax rather than vocabulary, or because they focus on specific constructs rather than more global measures. Nevertheless, the L1 effects that were found with these measures suggest that it may be possible for learners at roughly the same proficiency level in our sample to display different levels of lexical diversity, due to differences in their L1. For example, this might be because the learners with the different L1s are classified as having similar L2 proficiency overall, but some have higher lexical diversity whereas others have higher syntactic accuracy. This possibility is supported by our analyses, which show that there is very high variance in lexical diversity, even among learners with the same proficiency level who are engaging in the exact same writing task.

2 METHODOLOGY

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

This repository also contains the supplementary documents that we mention in the paper, including the main "Supplementary Information (SI)" document.

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 in the study's OSF repository (in the "Learner sample" folder).

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; 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) (Council of Europe, 2001). The curriculum is standardized, so learners with different L1s follow the same lessons and activities.

Each level in the EFCAMDAT consists of several lessons. After completing a lesson, learners are assigned a writing task (unique to that level) that they submit online, and then receive feedback on from a teacher.[2] These tasks (discussed in detail in the "Sample information" document), 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. However, they all include the same general format, whereby learners write text in response to a prompt following some lesson. Accordingly, the way in which we use the term "different writing tasks" (and later also "task effects") differs from how it is used in some contexts, though it aligns with some past research on tasks in the EFCAMDAT (e.g., Alexopoulou et al., 2015, 2017).

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.[3] 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), the structure of this dataset offers an opportunity to conduct our analyses on two similar but distinct learner samples, which serves as a form of replication.

Texts were randomly selected from this dataset, using stratified sampling so that the final sample is balanced across L1s, proficiency levels, and tasks, as outlined in the supplementary "Sample information" document (in the OSF repository). The final samples are listed in Table 1.

[2] Teachers also grade the texts that learners write; if a text receives a failing grade, then the learner simply repeats the lesson and/or task, until they do well enough to advance. Since learners usually pay for this service, they are motivated to use it in order to learn properly. In addition, although learners usually work on tasks alone at home, the large number of errors they make suggests that they do not usually use tools like spell-checkers.

[3] In the context of the present research, we use *topic* to refer to the general topic that each text revolves around (e.g., "Describing your favorite day") and *prompt* to refer to the prompt that learners wrote the text in response to (e.g., "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…"). We then use *task* to refer more broadly to every aspect of each writing task that can affect the lexical diversity of texts; this includes the topic being discussed and the specific prompt that learners responded to, as well as other potential factors, like the lesson that preceded that task. As we explain in more detail when presenting our approach to data analysis, our operationalization of *task* and *task effects* is therefore distinct from its operationalization in many other contexts (Alexopoulou et al., 2017), and we will not make a claim regarding which aspects of the tasks influenced learners' lexical diversity.

Table 1. Final learner samples.

L1s (nationalities) [a]	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) [b]
Tasks per corpus	95 (first) / 71 (second) [c]
Number of texts per L1 per task	10 [d]
Total number of texts (corpus)	8,500 (first) / 6,390 (second)
Total number of tokens (corpus)	701,990 (first) / 576,437 (second) [e]

[a] L1 in the EFCAMDAT is approximated based on learners' nationality, an approach that has been validated empirically (e.g., Alexopoulou et al., 2017); for more information, see the "Sample information" document.
[b] Every 3 EFCAMDAT levels correspond to a single CEFR level.
[c] There are 8 tasks per EFCAMDAT level in the first subcorpus, and 6 tasks per level in the second. One exception is task #51, in which texts from both subcorpora were classified under the first subcorpus due to limitations in the classification scheme, so this task was removed from both subcorpora.
[d] 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 shown in the "Sample information" document (in the OSF repository).
[e] This is based on data in the *wordcount* variable that is provided as part the EFCAMDAT Cleaned Subcorpus.

2.2 Lexical distance

We calculated lexical distance[4] between the L1s and English using *Swadesh lists* from the *Automated Similarity Judgment Program* (ASJP) (Wichmann et al., 2018), which are wordlists containing concepts that appear in nearly all languages, such as *water*, *full*, and *hear* (Swadesh, 1950). 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 et al., 2013; 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., 2008). 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

[4] See the supplementary information for an explanation why we use the term <u>*lexical*</u> *distance*.

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 the supplementary information. This includes research showing that it strongly correlates with expert cognancy judgments (Schepens et al., 2013), 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., 2013, 2020).

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 2. The wide range of distances from English highlight the diversity of the L1s in our sample. The median distances for most L1s is 1, reflecting a maximal distance between most words in the sample and their English translations. However, this does not pose an issue for our analyses, and the use of these lists has been validated despite this, as noted in the supplementary information. Furthermore, the resulting distances largely align 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, though each of these measures comes with its own limitations. For example, orthography-based LDN is more relevant to our written sample, but is problematic to use since we also examine L1s that use a different script than English (in which case the LDN is always maximal). Similarly, expert-based cognancy judgments can account for crosslinguistic relations between words in a more holistic way than LDN (e.g., by also considering morphological similarity), but they do not provide quantitative information on the degree of formal similarity across words, and are subjective. Likewise, psychotypological and psycholinguistic measures, which are based on perceived language distance, can more accurately measure the key variable of interest—crosslinguistic similarity as it is perceived by learners—but these measures depend on factors like the sample and methodology used to generate them, and are much harder to generate, especially for a range of diverse languages. All this is *not* to say that these measures are bad, as they can certainly be beneficial to use in various contexts. Rather, this is to give added context on why we decided to use phonological LDN as the key measure in our research, while acknowledging that other measures of similarity could be used too.

Nevertheless, as noted above, and as shown in detail in the supplementary information, this particular distance has been extensively validated, and is strongly correlated with these other distances. Accordingly, it is unlikely that the use of a different measure would substantially influence the results. This is supported by the particular findings of our study (especially the estimated marginal means in Table 4), which suggest that variation in the exact distances that were calculated would not substantially change our key findings. In addition, as discussed later, supplementary models with a binary measure of lexical distance (based on whether the L1 is Indo-European like English) replicated the results obtained when using Swadesh lists to measure distance.

Table 2. Lexical distance between each L1 and English.

	Lexical distance from English [c]				
L1[a,b]	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).

[a] L1s are ranked in increasing order of mean distance.

[b] These L1s were selected based on the availability of data in the EFCAMDAT Cleaned Subcorpus, as outlined in the "Sample information" document.

[c] Based on the phonological normalized Levenshtein distance from the closest synonym in the Swadesh lists. For information about the extensive validation of this measure, see the supplementary information.

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).

We assessed lexical diversity using the *measure of textual lexical diversity* (MTLD) (McCarthy, 2005). As described by McCarthy and Jarvis (2010, pp. 384–385), this is equivalent to the mean number of sequential words in a text that maintain a given TTR value (by default, .72). To calculate it, the TTR of the text is evaluated sequentially at every word; when the given TTR value is reached, the factor count increases by 1, and the TTR evaluation resets, as in the following example:

WORDA (TTR = 1.00) WORDB (1.00) WORDA (.67) ||FACTORS = FACTORS +1|| WORDA (1.00)…

For the final words in a text that do not form a full factor (the *remainder*), a partial factor is calculated, based on their TTR value. For example, a final TTR of .90 covers 36% of the range between 1.00 and the default .72, so it adds .36 to the factor count. Then, an MTLD value is calculated by dividing the number of words in the text by the total factor count; this is done in both forward and reverse processing of the text, and the mean of the two resulting values is the final MTLD.

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; Yan et al., 2020; Zenker & Kyle, 2021).

However, because the shorter the text the greater the contribution of the remainder to the MTLD, and because remainders represent approximations of lexical diversity, MTLD is less robust on short texts below a certain point (McCarthy & Jarvis, 2010). 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 50 words, this 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; 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.

2.4 Data analysis

We built mixed-effects linear regression models (Hox et al., 2018; Winter, 2019) for each corpus in our sample. Such models add *random effects* to traditional regression models—which contain only fixed effects—in order to account for grouping structures in the data, like multiple texts that were written by the same learner. This is important for controlling for dependencies between observations in the sample, and furthermore, the magnitude of these effects can be informative, as we will see when comparing task effects to the effects of L2 proficiency.

The models had 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 (fixed effects)**:
    a. *Lexical distance* (as a continuous variable), 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* (as a continuous variable), 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)[5]:
    a. *Learner*, to control for cases where learners had more than one text in the sample.[6]

[5] 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; see the supplementary information for details.

[6] Though most learners had only one text in the sample; see the “Sample information” document for details.

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.[7] 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 removed slightly under 5% of texts, which were classified as outliers based on extreme MTLD values, leaving 8,081 texts in the first corpus and 6,129 in the second; the rationale and process for this are presented in the "Outliers" section of supplementary information, which supports this removal but also shows that it does not substantially change our findings. Also, as further shown in the supplementary information, we checked the assumptions of the models and found no substantial issues (in terms of normality of residuals and random effects, homoscedasticity, lack of collinearity, and lack of influential observations), and compared our models with baseline models without lexical distance. Finally, to provide further insights, we also created several associated plots and tables, shown in the results.

[7] 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.

3 RESULTS

Figure 1 and Table 3 show that lexical diversity (MTLD) in L2 writing generally increases as learners' 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 imperfectly and with substantial variability.

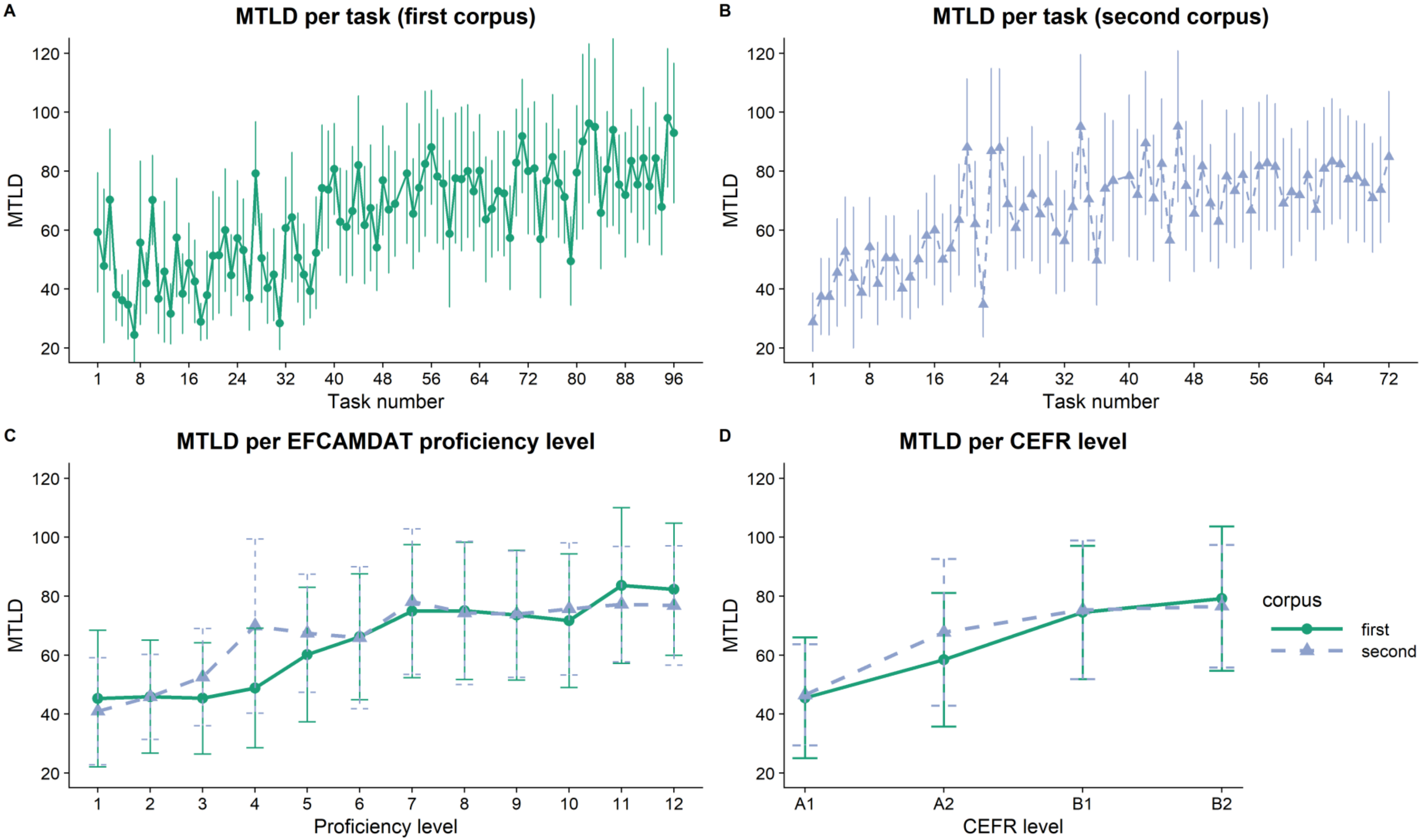

Figure 1. Mean lexical diversity (MTLD); error bars indicate one standard deviation. Listed per *task* in (A) and (B), per *EFCAMDAT proficiency level* in (C), and per *CEFR level* in (D). There are 8 tasks per EFCAMDAT proficiency level in the first corpus and 6 tasks per EFCAMDAT level in the second (each task appears only at a specific level; this is controlled for in the later mixed-effects models). There are 3 EFCAMDAT levels per CEFR level in both corpora. Raw correlations appear in the supplementary information.

Table 3. Mean lexical diversity (MTLD) per CEFR (L2 proficiency) level, corresponding to Figure 1D.

	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 2 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 3 show that lexical distance does *not* predict MTLD at any CEFR proficiency level (as they all have a non-significant R^2 that is lower than .001). Furthermore, Table 4 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.

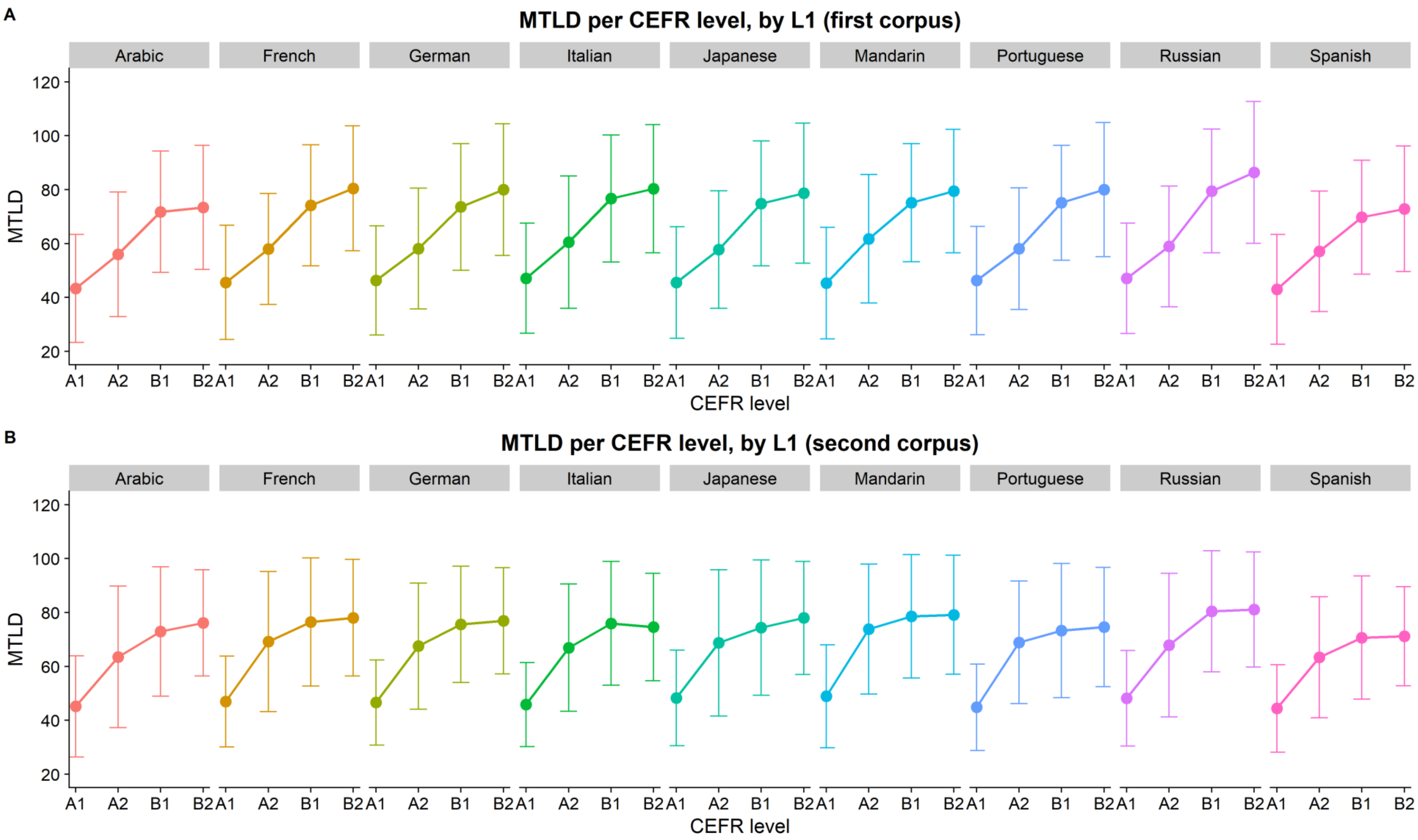

Figure 2. Mean lexical diversity (MTLD) per CEFR (L2 proficiency) level, by L1, in each corpus. Error bars denote one standard deviation. The exact values for these plots appear in the supplementary information.

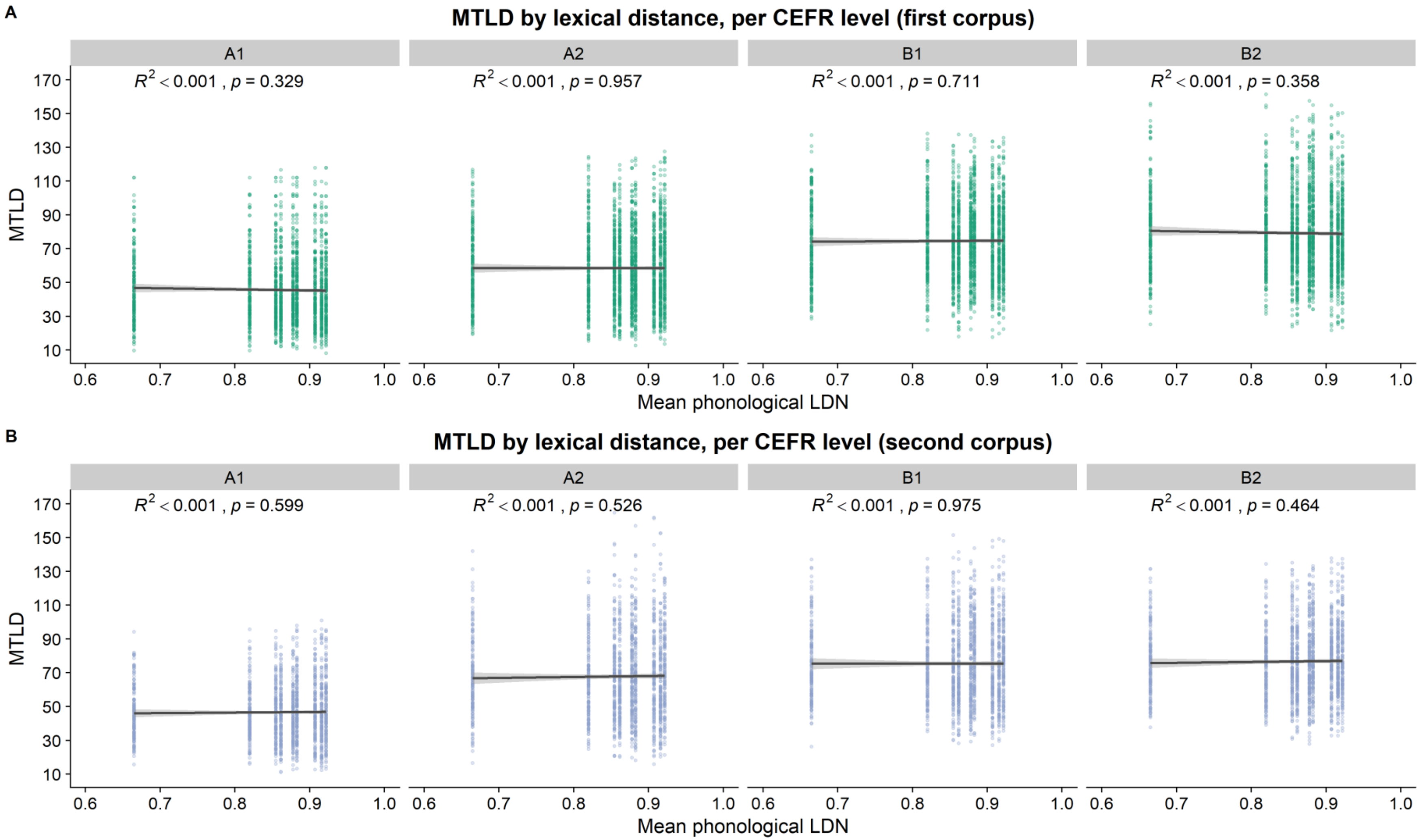

Figure 3. Scatterplots and linear models with lexical distance (LDN) as the predictor and lexical diversity (MTLD) as the response variable, per CEFR (L2 proficiency) level. Each model's R^2 and *p* appear in the corresponding panel. The grey bands around each line denote its 95% CI. Darker points denote 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 3 (range = 1,500–2,088).

Table 4. *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.

	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). There were 8,081 texts in the first corpus and 6,129 in the second.

[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 5 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.[8] 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.

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

Table 5. 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 [a]	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).

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 the key findings below, together with their implications for theory and practice.

4.1 The effect of crosslinguistic lexical similarity on lexical diversity

Our results show that there is no effect of crosslinguistic lexical similarity on written L2 lexical diversity, either in general or at low L2 proficiency levels. This was evident in the mixed-effects models (Table 5), 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 4), 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 3), 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 production, 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.[9] This supports the findings of Crossley and McNamara (2011), on the lack of effect of L1 on lexical diversity in task-based settings. 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,

[9] 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 4). 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 also contradict the facilitative effect of similarity on broad L2 proficiency (Schepens et al., 2013, 2020; van der Slik, 2010).

therefore, suggest that the effect of lexical similarity is 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, assessment, and research, educators, assessors, and researchers 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, as noted at the end of the introduction (§1), past studies on the EFCAMDAT found L1 effects for other linguistic phenomena, including accuracy of grammatical morphemes (Murakami, 2016), among various others (Alexopoulou et al., 2015; Jiang et al., 2014; Shatz, 2019), as did research on the Cambridge Learner Corpus (Murakami & Alexopoulou, 2016).

This suggests that L1 effects can be found in this sample, which is supported in the case of lexical diversity by the substantial variance we found in this measure, even within responses to the same task. Accordingly, the lack of similarity effect seems to likely be due to a difference between crosslinguistic influence on *lexical* rather than *functional* content, the former playing a more important role in meeting communicative demands of tasks. Nevertheless, it is important for future research to confirm these findings. Notably, such research can replicate the analyses on other learner samples, and especially ones where it is possible to track the rate of development of lexical diversity in specific learners over time. Such research can also examine a different target L2, and especially one that is not a lingua franca like English.

4.2 The effect of L2 proficiency on lexical diversity

As shown in the results (in Figures 1 and 2, and Tables 3 and 5), lexical diversity in writing 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 in the increase of lexical diversity, so they were able to use MTLD to distinguish significantly only between texts at the B1 and C1/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 1 and Table 3), 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 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. This is important for those who use lexical diversity to assess L2 proficiency, for example in language tests and L2 acquisition research.

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 1 and Table 5, a finding that 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 that is 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 §4.1. From a practical perspective, this highlights the importance of accounting for task effects when assessing lexical diversity, for example in language tests. In addition, this also highlights the importance of conducting further research on how different tasks influence lexical diversity, in order to better understand and control for task effects. Notably, such research could also attempt to disentangle how different aspects of tasks influence lexical diversity, for instance when it comes to the topic of tasks, their expected level of formality, and their educational context (e.g., preceding lesson). As we noted earlier (§2.1 and §2.4), this is *not* something that we did in the present research, since we examined task effects as a whole, by using mixed-effects models to analyze a tasks as a general grouping factor (i.e., random effect), without investigating any particular aspect of those tasks.

4.4 Conclusions

In our study, lexical similarity between learners' L1 and their target L2 did *not* increase or otherwise influence their lexical diversity in L2 English writing, 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, assessment, and research as they help understand and predict the patterns that appear in learners' L2 lexical diversity.

5 References

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

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

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

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

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

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/

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International corpus of learner English*. Universite Catholique de Louvain: Centre for English Corpus Linguistics.

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

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

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

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, 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

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. (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, 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

Kessler, M., Ma, W., & Solheim, I. (2022). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, *56*(4), 1163–1190. https://doi.org/10.1002/tesq.3096

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

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

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

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Lange eines Textes. *Zeitschrift Fur Literaturwissenschaft Und Linguistik*, *8*, 73–79.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. University of Memphis.

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

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.

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

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

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

Schepens, J., van der Slik, F., & van Houta, R. (2013). 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*, Article 104056.

https://doi.org/10.1016/j.cognition.2019.104056

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

Shatz, I. (2022). *The potential influence of crosslinguistic similarity on lexical transfer: Examining vocabulary use in L2 English* [University of Cambridge]. https://doi.org/10.17863/CAM.86440

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

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

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., 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

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

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

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/

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

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

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007

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

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