

Introducing and Evaluating a Measure of Lexical Diversity Across Word Classes

TAEHYEONG KIM  AND **TOVE LARSSON** 

*Northern Arizona University
Flagstaff, Arizona, USA*

HENRIK KAATARI 

*University of Gävle
Gävle, Sweden*

YING WANG 

*Karlstad University
Karlstad, Sweden*

PIA SUNDQVIST 

*University of Oslo
Oslo, Norway*

Abstract

Lexical diversity (LD) has been shown to be a strong predictor of second language (L2) proficiency. However, most current indices combine all word classes into a single measure and thus only capture the broadest patterns of lexical variation. In addition, researchers interested in examining usage patterns across word classes currently lack access to measures of LD that are both more linguistically interpretable *and* robust to text length. In response to these issues, this paper introduces a methodology for examining part-of-speech (POS)-specific LD indices (e.g., verb and noun diversity) that apply the moving-average type-token ratio (MATTR), a measure validated for its stability across text lengths and reliability in shorter texts. We also evaluate these measures by comparing them to traditional, so-called *omnibus* LD measures for interpreting L2 corpus data. The results show that examining LD within different parts of speech can help researchers disentangle the broader developmental trend observed with omnibus measures. These findings highlight how POS-specific

LD provides more linguistically interpretable description of L2 lexical development, thus complementing traditional omnibus measures.

doi: 10.1002/tesq.70154

INTRODUCTION

The concept of lexical complexity has repeatedly been found to be associated with second language (L2) proficiency in recent research (e.g., Yoon, 2017). As a sub-construct of lexical complexity, lexical diversity (LD) captures the extent to which writers or speakers use distinctive words (McCarthy & Jarvis, 2010). LD is typically operationalized through quantitative metrics that measure the proportion of distinctive words (i.e., types) to the total number of words (i.e., tokens) in L2 production.

In the field of second language acquisition (SLA), a substantial body of research has shown that LD is a strong predictor of L2 written and oral proficiency levels, as measured by standardized tests, proficiency scales, and/or the rated quality of learners' written and oral production (e.g., Bulté & Roothoof, 2020; Zhang, 2022). However, as outlined in Biber, Gray, Staples, and Egbert (2020); Larsson and Biber (2024), studies that address linguistic questions must ultimately return to the linguistic interpretation of their findings. Linguistic interpretability—defined as the extent to which “scale and values [of linguistic variables] represent a real-world language phenomenon that can be understood and explained” (Egbert, Larsson, & Biber, 2020, p. 24)—is a central concern for all such studies. In LD research, for instance, most current indices combine all word classes in an omnibus category. Although omnibus LD measures represent variation in lexical choice and have been shown to account for differences in L2 proficiency, they capture only aggregated patterns of lexical variation, and their linguistic interpretability is limited if the researchers seek to explain whether or not different aspects of lexis drive such variation. Against this backdrop, this study explores part-of-speech (POS)-specific LD as a complementary measure that offers a means of enhancing the linguistic interpretability of existing LD indices.

To illustrate how studying the POS-specific diversity of lexis can be helpful for more detailed linguistic interpretation, we focus on two POS categories: verbs and nouns. These grammatical categories have attracted particular attention in SLA because of their distinct discourse functions. Lexical verbs are key to expressing functions such as action and personal stance. However, studies have shown that L2 writers tend

to rely on a narrow set of verbs and repeat many “conversational verbs” (e.g., *think*) that are typical of informal speech (Granger & Paquot, 2015). Nouns, by contrast, are central to elaborating on topics and information. Prior studies have consistently shown that nouns are strongly associated with the literate-style registers, such as academic registers (Biber, Johansson, Leech, Conrad, & Finegan, 2021, p. 237).

Importantly, verbs and nouns have been shown to display distinct L2 developmental patterns for sub-constructs of lexical complexity. For example, Durrant, Dirdal, and Tveitan (2025) reported “divergent patterns for different parts of speech” (pp. 34, 35) for lexical sophistication in L2 English writing by secondary school students, and hypothesized that examining different POS categories can help explain inconsistent results from previous studies that combined different POS categories together. In terms of LD, Paquot (2019) observed a non-significant but systematic increase across proficiency levels in L2 writing for verb diversity, but not for noun diversity. Building on this line of inquiry, Nasser and Thompson (2021) investigated LD in dissertation abstracts produced by English first language (L1) and L2 postgraduate students. Although they found that verb diversity and noun diversity did not show significant between-group differences, they reported a weak correlation ($r = .4$) between omnibus LD and noun diversity, leading the authors to conclude that these “should indeed be regarded as separate categories” (p. 7).

However, apart from the studies mentioned above, there has been very limited attention paid to POS-specific LD, which is unfortunate given the differing developmental trends reported in prior research for different POS categories. As an added complication, the existing studies have so far largely relied on formulas that do not adequately control for text length effect, such as the indices implemented in the Lexical Complexity Analyzer (LCA; Lu, 2012). The LCA measures use traditional type-token ratio (TTR) and its variants. The reason that these measures are potentially problematic is that previous validation research has consistently shown that TTR and its mathematical transformations (e.g., Root TTR) are not only highly sensitive to text length but also unreliable for shorter texts (Koizumi & In’nami, 2012; Zenker & Kyle, 2021). For example, based on the formula, we can expect a priori that longer texts will have less diversity, just because they are longer (i.e., the likelihood of repeating words increases with text length). It should also be noted that mathematical transformations change the nature of the data, often to a scale that is difficult to interpret linguistically.

In sum, we note two issues with prior applications of LD measures: (a) the most commonly used measures do not take the POS of a word into consideration, thus potentially combining words with different

developmental trajectories, and (b) available measures that do take the POS into consideration do not account for well-known issues related to text length.

In response to these issues, this paper introduces a technique for studying LD within different POS categories while applying the moving-average type-token ratio (MATTR; Covington & McFall, 2010), validated for its robustness to text length and reliability in shorter texts. We also provide code that enables researchers to apply this approach to their own studies. In addition, we empirically evaluate this technique guided by the following three research questions:

1. To what extent are omnibus LD and POS-specific LD associated, such that if we see high/low omnibus LD, we also see high/low noun and verb diversity?
2. To what extent do the developmental trajectories of omnibus LD vs. POS-specific LD differ across grade levels in L2 writing?
3. To what extent do POS-specific LD indices explain additional variance in developmental differences beyond omnibus LD?

The subcorpus used in this study comes from the Swedish Learner English Corpus (SLEC; Kaatari, Wang, & Larsson, 2024), and it encompasses 400 L2 English argumentative essays written by Swedish junior and senior high school students. The essays were manually cleaned of spelling errors to avoid inflating type counts. As LD may differ across topics or prompts, this corpus is helpful for our validation study in that all essays were written on the same topic (*How to lead a good life*). The essays are evenly distributed across grade levels, with 100 essays from each grade (grades 8 through 11).

In terms of the LD measure, we used MATTR¹ (Covington & McFall, 2010), which is designed to reduce the well-known problem that TTR declines as text length increases. It works by sliding a fixed-size window of tokens across the text, calculating the TTR within each window, and then averaging those values. As a first step of the analysis, we had to decide on a window size, given that the window size (which equals the minimum token count) of MATTR is an important parameter that affects not only the resulting values but also their interpretation (Bestgen, 2024; Covington & McFall, 2010). As POS analyses might require a smaller window size than omnibus analyses, we tested how stable the measure remains with smaller window sizes than have previously been validated (e.g., Zenker & Kyle, 2021).

¹ As an initial step in applying the existing LD approach to POS-specific measures, the present report used MATTR, but it should be noted that there are other LD indices (e.g., HD-D, MTL) that have been shown to be robust to different text length (e.g., Bestgen, 2024) and could likewise be adapted for POS-specific analyses. Evaluating these alternatives represents an important direction for future research.

Previous studies have typically used a window size of 50 tokens when calculating MATTR, which in turn required texts of at least that length (e.g., Koizumi & In'nami, 2012; Zenker & Kyle, 2021). In contrast, this study examines how MATTR behaves with smaller window sizes (10, 20, and 30 tokens) for the POS-specific analyses of LD. To assess the stability of MATTR with smaller window sizes, we excluded texts containing fewer than 40 lexical verb tokens or fewer than 40 noun tokens. This threshold allows sufficient range to observe how MATTR values stabilize as token counts increase beyond each window size, while also retaining a larger number of texts for POS-specific analysis. For each remaining text, we extracted the first 40 tokens for each word class to ensure comparability across samples.

MATTR values were then computed using window sizes of 10, 20, and 30. For each window size, MATTR was first calculated on the smallest possible segment (i.e., the first 10, 20, or 30 tokens, corresponding to the respective window size). Following the standard sliding-window procedure (Covington & McFall, 2010), the number of tokens included in the analysis was then increased incrementally by one (e.g., from 10 to 11 tokens for a window size of 10 and from 20 to 21 tokens for a window size of 20), up to the full set of 40 tokens. In this way, the MATTR values of each window size (10, 20, and 30) were evaluated across progressively longer token sequences (i.e., increasing text length), allowing us to assess how stable MATTR values remain as text length increases while the window size is held constant.

Stability was evaluated using two criteria: (a) variability across token counts should be minimal, operationalized as overlapping 95% confidence intervals (CIs) of all mean MATTR values for each grade level across token counts, and (b) the relative ordering of grade levels in MATTR should remain stable across token counts, operationalized as the absence of intersections among grade-level trajectories. Considering both criteria, a window size of 30 was the only size for which (a) the 95% CIs of all MATTR means overlapped for each grade level and (b) no intersections of the grade-level trajectories were observed for both noun and verb diversity (see Data S1 for figures and interpretation).

We therefore proceeded with MATTR using a window size of 30 tokens and a sliding increment of one token in the subsequent analyses. We included only texts containing at least 30 lexical verbs and 30 nouns that met this minimum token threshold. This resulted in a final data set of 347 texts (from the initial 400) for the subsequent analyses. The remainder of this brief report is structured as follows: The “[Introducing Lexical Diversity Indices](#)” section introduces the new measures. The “[Evaluation](#)” section provides an overview of the results

from the evaluation, and the “Conclusion” section concludes the report.

INTRODUCING LEXICAL DIVERSITY INDICES

To compute POS-specific and omnibus LD values, the first author developed a custom Python script, which is openly available as the posLD package (Kim, 2026). Instructions and updates are available at <https://github.com/taehyeongterrykim/posLD>. All texts were processed using spaCy (Honnibal & Montani, 2025). Texts were tokenized and POS-tagged following the Universal Dependencies tag set using the *en_core_web_trf* pipeline. According to spaCy’s reported benchmarks, this model achieves approximately 98% POS tagging accuracy (spaCy, n.d.). To verify tagging accuracy in the present data set, we manually inspected two random samples of 100 tokens each that had been automatically tagged as verbs (VERB) and nouns (NOUN), respectively. No tagging errors were observed in either sample. After tagging, the tokens were subsequently filtered into lexical categories (i.e., lexical verbs, nouns, content words, and all words) and then used as input to compute MATTR. This study specifically focuses on verbs and nouns, but the diversity of all words and all content words, respectively, were also computed for comparison. In sum, we computed:

- MATTR_{30_verb} : The moving average type-token ratio of a segment length of 30 lexical verbs (excluding auxiliaries) in a text.
- MATTR_{30_noun} : The moving average type-token ratio of a segment length of 30 nouns (excluding pronouns and proper nouns) in a text.
- MATTR_{30_cw} : The moving average type-token ratio of a segment length of 30 content words in a text.
- MATTR_{30_aw} : The moving average type-token ratio of a segment length of 30 words in a text.

EVALUATION

Correlations Between POS-Specific and Omnibus LD Indices

To address our first research question, Pearson’s correlations were computed among all indices to investigate trends in the POS-based LD measures (noun only and verb only) vs. omnibus LD measures (all words and all content words, respectively) in our texts. Conceptually,

TABLE 1

Pearson's Correlation Matrix Among Omnibus (aw, All Words; cw, Content Words) and POS-Specific (Verb vs. Noun) MATTR Measures

	MATTR _{30_aw}	MATTR _{30_cw}	MATTR _{30_verb}
MATTR _{30_cw}	.81		
MATTR _{30_verb}	.65	.76	
MATTR _{30_noun}	.58	.77	.48

the POS-specific measures are nested within the broader construct of LD, and therefore, some degree of correlation between these indices is expected. The results (Table 1) show moderate-to-strong correlations between the POS-specific and omnibus measures, indicating that they capture related subcomponents of lexical diversity. At the same time, the correlations remain below commonly used thresholds for multicollinearity (e.g., .80–.90; Plonsky & Ghanbar, 2018), suggesting that the measures are not redundant and may provide more fine-grained information within the broader LD construct.

In other words, a text with high omnibus LD does not necessarily exhibit equally high verb or noun diversity. POS-specific measures therefore allow researchers to examine which word classes contribute to overall LD patterns in L2 writing. This finding is noteworthy given that lexical verbs and nouns together constitute a substantial proportion of all the lexical items in the texts. On average per text in our corpus, lexical verbs alone accounted for 32% of content words and nouns alone accounted for 36%. Despite their substantial contribution to the overall lexical pool, diversity within these word classes does not fully mirror the patterns observed in omnibus LD measures. The correlation between verb and noun diversity itself was also moderate ($r = .48$), suggesting that the diversity of these two word classes varies somewhat independently across texts. Taken together, these results indicate that POS-specific LD measures can provide a more fine-grained perspective on LD diversity patterns in L2 writing that complements omnibus LD indices, as opposed to providing redundant information.

Example (1) illustrates how POS-based LD indices can describe linguistic differences hidden in the omnibus measures, with verbs shown in bold and nouns underlined. In this case, the writer's MATTR values show low verb diversity (MATTR_{30_verb} = .35) but much higher noun diversity (MATTR_{30_noun} = .61). The verb lemma *have*, shown in red, is densely repeated alongside a narrow set of lexical verbs, while a much wider range of nouns is used, most of which occur just once except for *life*. In this case, the broader content-word

(MATTR_{30_cw} = .63) and all-word diversity (MATTR_{30_aw} = .71) mask the writer's more fine-grained patterns of diversity in different word classes.

(1) I **believe** that a good life **makes** you **feel** good inside, if you **shopping** with your friends or if you **have** a family and a person in your life as really in love with you and you **have** a really healthy relationship. Then you **have** a good life if you are happy to **have** no bad presences. All the things you **do** and **makes** you happy **doing** so you **have** a good health and good life. But a good life can **have** bad things like you **have** not so much money and you cannot **buy** some much things like clothes and makeup and other stuff. If you **had** a bad education and you don't **have** a job. You can also **have** not so good health you **eat** unhealthy food and you be fat. It is my opinion.

(H_8_F_24_5_C)

Developmental Trajectories of LD Indices

To examine our second research question, the developmental trajectories of different LD indices, we first visualized each LD measure in terms of mean MATTR values and their 95% confidence intervals across the grade levels. Then, we conducted a series of one-way ANOVAs with grade level as the independent variable and each LD measure as the dependent variable, followed by Tukey post hoc comparisons (see Data S2 for full ANOVA and Tukey post hoc results).

As shown in Figure 1, although the overall trend is similar for noun and verb diversity, different fine-grained developmental trajectories emerge. The noun diversity is higher than verb diversity in Grades 8–10; however, by Grade 11, verb diversity catches up with noun diversity. Importantly, the increase in verb diversity from Grade 10 to 11 is marked and statistically significant ($p < .001$) with a medium-to-large effect size ($d = .58$; Cohen, 1988). By contrast, the corresponding change in noun diversity is smaller and not significant ($p = .18$, $d = .30$). The POS-specific trajectories therefore suggest that the gain in omnibus LD from grade 10 to 11 is driven primarily by verb diversity, rather than by noun diversity. We also note that the LD measure for all words fails to capture this marked increase from Grade 10 to 11.

Examples (2 and 3) illustrate such developmental differences in verb diversity, with all verbs underlined, and repeated verbs additionally color-coded. Example (2) is a truncated example from a grade 11 student who shows very high verb diversity (MATTR_{30_verb} = .97),

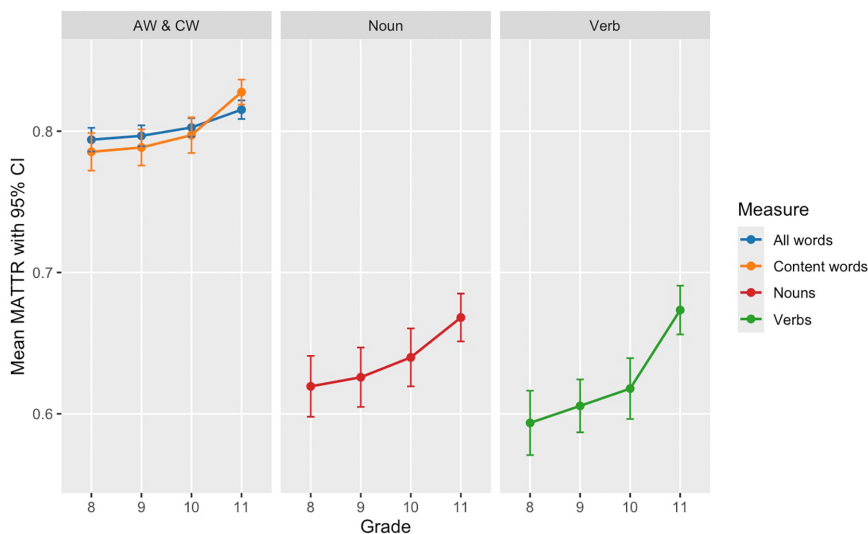


FIGURE 1. Developmental trajectories of lexical diversity across grade levels.

whereas Example (3) is from a Grade 10 student who shows low verb diversity ($\text{MATTR}_{30_verb} = .48$).

(2) Both physically and mentally, a poorly functioning body directly **hinders** you from **achieving** a good life, because it **lessens** your ability to **act** on your own will. Naturally it is still possible to **achieve** happiness anyway, poor health only **makes** it more difficult to do so.
(G_2_S_M_22_82_C)

(3) It can **affect** your mood in both negative and positive ways. You should always **feel** that you **have** someone to **talk** to if you **feel** bad about something or just want to **talk**. You should always **feel** safe and secure with the family because it is usually those you are around most of the time.
(G_1_Y_F_21_113_C)

Explanatory Power of POS-Specific LD Indices

To examine whether POS-specific LD indices account for additional variance in grade level beyond omnibus LD, ordinal regression models were fitted predicting grade level. Baseline models included either all-word MATTR or content-word MATTR, and extended models additionally included verb and noun diversity. Likelihood-ratio tests were

TABLE 2
Model Comparison: All-Word MATTR

Model	Predictors	logLik	AIC	χ^2	df	<i>p</i>
Model 1	mattr _{30_aw}	-470.79	949.57			
Model 2	mattr _{30_aw} + mattr _{30_verb} + mattr _{30_noun}	-463.53	941.07	12.51	2	.002

used to assess whether the inclusion of POS-specific measures improved model fit.

Likelihood-ratio tests indicated that including the POS-specific measures significantly improved model fit for both the all-word model ($\chi^2(2) = 12.51$, $p = .002$; see Table 2) and the content-word model ($\chi^2(2) = 7.04$, $p = .029$; see Table 3).

To further examine the contribution of individual predictors, the final models are presented in Tables 4 and 5. In both models, verb diversity emerged as a significant predictor of grade level (all-word model: $\beta = 4.41$, $SE = 1.33$, $p < .001$; content-word model: $\beta = 4.14$, $SE = 1.59$, $p = .009$), whereas noun diversity and the omnibus MATTR measures were not significant predictors. The results suggest that developmental differences in LD are not evenly distributed across word classes. The lack of significance for omnibus MATTR and noun diversity in the final models further indicates that aggregating across word classes may obscure important fine-grained variation, potentially diluting differences across proficiency levels that are more clearly captured by specific lexical categories such as verbs.

TABLE 3
Model Comparison: Content-Word MATTR

Model	Predictors	logLik	AIC	χ^2	df	<i>p</i>
Model 1	mattr _{30_cw}	-468.06	944.12			
Model 2	mattr _{30_cw} + mattr _{30_verb} + mattr _{30_noun}	-464.54	941.07	7.04	2	.029

TABLE 4
Final Ordinal Regression Model (All-Word and POS)

Predictor	β	SE	<i>z</i>	<i>p</i>
mattr _{30_aw}	2.75	4.13	0.67	.506
mattr _{30_verb}	4.41	1.33	3.31	<.001
mattr _{30_noun}	0.75	1.28	0.59	.555

TABLE 5
Final Ordinal Regression Model (Content-Word and POS)

Predictor	β	SE	z	p
mattr _{30_cw}	2.38	3.61	0.66	.509
mattr _{30_verb}	4.14	1.59	2.60	.009
mattr _{30_noun}	0.34	1.65	0.20	.839

CONCLUSION

The present study introduced and evaluated POS-specific LD indices that apply the MATTR approach. The results showed that a window size of 30 tokens produces stable MATTR values across different token counts for both verbs and nouns, and that examining LD within individual POS categories can provide a more fine-grained perspective on lexical variation in L2 writing. At the same time, the findings should be interpreted in light of the study's scope. The analyses were based on texts from a single genre, topic, and learner population, which was an intentional methodological decision to control for contextual variation. Nevertheless, this design limits the generalizability of the findings, and future research should examine whether similar patterns emerge across different contexts. We hope that this approach and the accompanying code will be helpful for researchers seeking a more fine-grained understanding of lexical variation and its development in L2 production.

FUNDING

This work was funded by Riksbankens Jubileumsfond (Reference No. P23-0437).

CONFLICT OF INTEREST STATEMENT

There are no relevant financial or non-financial competing interests to report.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

THE AUTHORS

Taehyeong Kim is a PhD candidate in Applied Linguistics at Northern Arizona University. His research interests include corpus linguistics and quantitative research methods, specifically in relation to how these are applied to topics pertaining to second language acquisition, linguistic complexity, and register variation.

Tove Larsson is an Assistant Professor of Applied Linguistics at Northern Arizona University. She specializes in research methods and corpus-based analyses of grammatical complexity and L2 writing. Her work appears in journals such as *Corpora*, *Language Teaching*, *Research Methods in Applied Linguistics*, and *TESOL Quarterly*. She is on the editorial boards of journals such as *Journal of Second Language Writing* and *Journal of English for Academic Purposes*.

Henrik Kaatari is Senior Lecturer in English Linguistics at the University of Gävle, Sweden. His research interests are in the field of learner corpus research with a focus on syntactic/grammatical variation and linguistic complexity.

Ying Wang is Associate Professor of English linguistics at Karlstad University, Sweden. She specializes in corpus-based research on L2 writing and language use in academic and public contexts, with a focus on phraseology. She is the author of *The Idiom Principle and L1 Influence* (John Benjamins, 2016) and her recent work appears in journals such as *Studies in Second Language Acquisition*, *Journal of English for Academic Purposes*, and *English for Specific Purposes*.

Pia Sundqvist is Professor of English Education at the University of Oslo. She specializes in informal language learning and the assessment of L2 oral proficiency. She has published extensively and is the author of *Extramural English in Teaching and Learning* (with Sylvén, Palgrave Macmillan, 2016) and *Testing Talk* (with Sandlund, Bloomsbury Academic, 2024). She is an associate editor with *Innovation in Language Learning and Teaching* and *Language Learning and Technology* and the past president of the Swedish Association of Applied Linguistics.

REFERENCES

- Bestgen, Y. (2024). Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, 74(3), 638–671. <https://doi.org/10.1111/lang.12630>
- Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46, 100869. <https://doi.org/10.1016/j.jjeap.2020.100869>
- Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of spoken and written English*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.232>
- Bulté, B., & Roothoof, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, 91, 102246. <https://doi.org/10.1016/j.system.2020.102246>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

- 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>
- Durrant, P., Dirdal, H., & Tveitan, V. D. (2025). Vocabulary sophistication in children's L2 school writing. *International Journal of Learner Corpus Research*, 11(1), 17–46. <https://doi.org/10.1075/ijlcr.23025.dur>
- Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Press. <https://doi.org/10.1017/9781108888790>
- Granger, S., & Paquot, M. (2015). Lexical verbs in academic discourse: A corpus-driven study of learner use. In H. Basturkmen (Ed.), *English for academic purposes* (pp. 193–214). Abingdon, Oxon: Routledge.
- Honnibal, M., & Montani, I. (2025). *spaCy (Version 3.8.4) [Computer software]*. Explosion. Retrieved from <https://spacy.io>
- Kaatari, H., Wang, Y., & Larsson, T. (2024). Introducing the Swedish learner English corpus: A corpus that enables investigations of the impact of extramural activities on L2 writing. *Corpora*, 19(1), 17–30. <https://doi.org/10.3366/cor.2024.0296>
- Kim, T. (2026). *posLD (Version 0.1.1) [Computer software]*. Python Package Index. Retrieved from <https://pypi.org/project/posLD/>
- Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4), 554–564. <https://doi.org/10.1016/j.system.2012.10.012>
- Larsson, T., & Biber, D. (2024). On the perils of linguistically opaque measures and methods: Toward increased transparency and linguistic interpretability. In P. Crosthwaite (Ed.), *Corpora for language learning* (pp. 131–141). London: Routledge.
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208. <https://doi.org/10.1111/j.1540-4781.2011.01232.x>
- McCarthy, P. M., & Jarvis, S. (2010). MTL, 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>
- Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47, 100511. <https://doi.org/10.1016/j.asw.2020.100511>
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145. <https://doi.org/10.1177/0267658317694221>
- Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R² values. *The Modern Language Journal*, 102(4), 713–731. <https://doi.org/10.1111/modl.12509>
- spaCy. (n.d.). *English: Available trained pipelines for English*. Retrieved from <https://spacy.io/models/en>
- 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>
- 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, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics*, 32(3), 371–396. <https://doi.org/10.1111/ijal.12420>

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. The stability of different MATTR window sizes.

Data S2. One-way ANOVA results and Tukey post hoc comparisons for “Developmental Trajectories of LD Indices” section.