THE INTERPLAY BETWEEN SYNTACTIC AND LEXICAL COMPLEXITY

Souvztažnost mezi lexikální a syntaktickou složitostí

Christopher WILLIAMS

https://doi.org/10.5817/CZ.MUNI.P280-0534-2023-11

Abstract: *This article reports on research into the relationship between syntactic and lexical complexity, the aim being to discover whether the presence of one measure correlates with the presence of another. 100 argumentative essays written by L1-Czech L2-English students were analysed using the LSA and L2SCA. The results indicate that use of complex nominals and clause length, both features of syntactic complexity, have a strong positive correlation with lexical complexity measures.*

Key words: *syntactic complexity, lexical complexity, LSA, L2SCA, essay writing*

Abstrakt: *Tento článek podává zprávu o výzkumu vztahu mezi syntaktickou a lexikální složitostí a jeho cílem je zjistit, zda přítomnost jedné míry koreluje s přítomností druhé. Pomocí LSA a L2SCA bylo analyzováno 100 argumentačních esejů napsaných českými studenty angličtiny. Výsledky naznačují, že použití komplexních substantiv a délka vět, což jsou dva rysy syntaktické složitosti, mají silnou pozitivní korelaci s mírami lexikální složitosti.*

Klíčová slova: *syntaktická složitost, lexikální složitost, LSA, L2SCA, psaní eseje*

1 Introduction

Research in second language (L2) writing proficiency has centred on the construct of complexity, accuracy, and fluency (CAF), widely recognised as key indicators in oral and written assessments of language learners (Housen & Kuiken, 2009). Both syntactic complexity (SC) and lexical complexity (LC) are multifaceted components of the complexity branch of the CAF framework. SC involves aspects such as sentence length, subordination, and coordination, while LC covers lexical density, diversity, and sophistication.

Studies have shown that complexity measurements can be an indication of writing proficiency. Learners that write using longer sentences or make use of less common vocabulary (Crossley et al., 2010; McNamara et al. 2013) tend to score higher in essay writing assignments.

2 Syntactic complexity

Though lacking an agreed upon definition, syntactic complexity (SC) is generally understood as referring to the range and sophistication of language being used (Lu, 2010), with more sophisticated structures being more present in the language production of more proficient learners. It is important to note that SC can be affected by, amongst others, the learner's L1 (Lu & Ai, 2015) and the task they are completing (Qin & Uccelli, 2016).

The L2 syntactic complexity analyser (L2SCA) (Lu, 2010) uses 14 indices to evaluate the complexity of a text (see Table 1 for an overview). Seven of the indices are based on the t-unit, which has been defined as the "shortest grammatically allowable sentences into which writing can be split or minimally terminable unit" (Hunt, 1965, p. 20). In other words, a t-unit is a main clause plus all of its dependent clauses.

Table 1
Syntactic complexity indices, labels, and formulae as used in the L2SCA (Lu, 2010)

Area	Label	Description
Length of production unit	MLC	Mean length of clause
	MLS	Mean length of sentence
	MLT	Mean length of T-unit
Amount of subordination	C/T	Clauses per T-unit
	CT/T	Complex T-unit
	DC/C	Dependent clause per clause
	DC/T	Dependent clause per T-unit
Amount of coordination	CP/C	Coordinate phrases per clause
	CP/T	Coordinate phrases per T-unit
	T/S	T-units per sentence
Degree of phrasal sophistication	CN/C	Complex nominals per clause
	CN/T	Complex nominals per T-unit
	VP/T	Verb phrases per T-unit
Overall sentence complexity	C/S	Clauses per sentence

Length-based measures have been used with some degree of success to indicate the proficiency of L2 writing. Mean length of clause (MLC) has been associated with the quality of argumentative essays (see Chen et al., 2014; Kim, 2014; Li, 2015; Qin & Uccelli, 2016; Williams, 2023), as has mean length of t-unit (MLT) (see Casal & Lee, 2019; Kim, 2014; Yang et al., 2015) and

mean length of sentence (MLS) (see Chen et al., 2014; Taguchi et al., 2013). It should be noted that length-based measures are limited in their application, as though they have some use at broadly determining proficiency categories, they are not so discerning when it comes to the more subtle distinctions between adjacent ability bands (for example, distinguishing between B1 and B2). This issue becomes more evident the higher up the ability bands one moves (Paquot, 2018), as sentence length tends to plateau once learners are able to express themselves through, for example, phrasal complexity.

In addition to length-based measures, Kim (2014) also identified complex t-units per t-unit (CT/T) and complex nominals per clause (CN/C) as strong indicators of proficiency in essay writing. The use of complex nominals per clause or t-unit has often been found to correlate with the quality of more academically focused writing tasks (see Khushik & Huta, 2020; Larsson & Kaatari, 2020; Williams, 2023).

3 Lexical complexity

Lexical complexity research has primarily focused on measures of density, diversity, and sophistication. Lexical density is measured as the ratio of content words (nouns, verbs, adjectives, and adverbs) to the total number of words in a text. Diversity measures (number of different words, type-token ratios, etc.) have consistently found a positive relationship with L2 writing quality. With regards to sophistication, Johnson et al. (2013; 2016) found that the use of low-frequency vocabulary was an indication of L2 performance. A word is considered sophisticated if it does not appear on the British National Corpus' list of the 2000 most frequent words. Please see Appendix for a full list of the lexical complexity measurements and their formulae as used in this study.

4 Methodology

The aim of this study is to investigate the relationship between the 14 syntactic and 23 lexical complexity measurements as used in the L2SCA and LCA, with the intention of identifying a correlation between measurement pairs. Should a strong correlation exist between a measure of syntactic complexity and lexical complexity, it may be that that measure alone can be used as a reliable indicator of proficiency when applied to argumentative essay writing. This in turn would assist with simplifying the tasks of preparing lessons, giving feedback, or assessing student work, as the teacher would be able to concentrate attention on such a measure.

Little research appears to have been conducted on this relationship. Kovacevic (2018), in a similar study, found a large number of significant correlations between syntactic and lexical complexity measures, though concluded that they could only partly account for each other.

Based on the given objectives, the following research question guided the study:

What correlations exist between lexical and syntactic complexity measurements?

Should a strong correlation be found, further investigation will be necessary to determine the efficacy of the measurement's application. After all, correlation does not always equal causation.

4.1 Data set

This study is based on the dataset collected as part of my doctoral research (see Williams, 2023 for a more detailed description), in which I analysed the syntactic and lexical complexity features of 100 argumentative essays written by L1-Czech L2-English students in their final year of high school. Each participating student completed the same task in the same conditions, writing by hand their response to the statement *Some people think that teachers should be paid according to how much their students learn* within 45 minutes and using between 160–180 words. The essays were marked by two raters according to task fulfilment, grammar, and vocabulary, and then processed for syntactic complexity using the L2SCA (Lu, 2010) and lexical complexity using the LCA (Lu, 2012). The resulting complexity measurements from the Williams (2023) study were then analysed using Spearman's rank correlation.

5 Results

The data analysis revealed 101 significant correlations between syntactic and lexical complexity pairings. In this section, these correlations will be presented based on the category of syntactic complexity: length-based, subordination, coordination, phrasal, and sentence.

Table 2 presents the length-based complexity measurements (MLS, MLT, MLC) as they correlate to lexical complexity. There was no correlation found between these measures and verb sophistication (VS1, VS2) or verb variation (SVV1), hence their omission from the table. MLS has a weak to moderate correlation with the number of different words found in the first 50 (NDWZ-50), suggesting that as sentence length increases, so does the likelihood of unique words. A weak negative correlation exists between MLS and adverb variation (ADVV), the only correlation with ADVV, indicating that MLS increases, the occurrence of unique adverbs decreases. This could suggest that the participants that wrote longer sentences reused adverbs rather than introducing new ones. As MLT increased, so did lexical diversity measures (LD, NDWZ-50, NDW-ER50, NDW-ES50, MSTTR, CTTR, RTTR, LOGTTR, and UBER).

MLC correlates with all LC indices but NDWZ-50 and ADVV. The strongest correlations are with diversity measurements (CTTR and MTTR); however, strong correlations exist in all lexical complexity dimension (density, diversity, and sophistication), indicating that MLC may be used as a single measure to capture text complexity.

Table 2
Correlation between lexical complexity and length-based syntactic complexity measurements

LC Indices	**SC length of production unit indices**					
	MLS	Rs *(p)*	MLT	Rs *(p)*	MLC	Rs *(p)*
LD			0.231 (0.020)		0.242 (0.015)	
LS1		-		-	0.342 (<0.0001)	
LS2		-		-	0.302 (0.002)	
CVV1		-		-	0.341 (<0.0001)	
SVV2		-		-	0.340 (<0.0001)	
NDW		-		-	0.363 (<0.0001)	
NDWZ-50	0.267 (0.007)		0.280 (0.005)		-	
NDW-ER50		-	0.286 (0.004)		0.317 (0.001)	
NDW-ES50			0.255 (0.010)		0.296 (0.003)	
TTR		-		-	0.360 (<0.0001)	
MSTTR		-	0.234 (0.012)		0.329 (<0.0001)	
CTTR		-	0.232 (0.020)		0.441 (<0.0001)	
RTTR		-	0.230 (0.020)		0.441 (<0.0001)	
LOGTTR		-	0.205 (0.041)		0.387 (<0.0001)	
UBER		-	0.218 (0.029)		0.436 (<0.0001)	
LV		-		-	0.382 (<0.0001)	
NV		-		-	0.256 (0.010)	
VV1		-		-	0.392 (<0.0001)	
ADJV		-		-	0.219 (0.023)	
ADVV	-0.200 (0.046)			-		-

The subordination measures and corresponding lexical complexity correlations are shown in Table 3. Measures of subordination and overall sentence complexity (Table 6) had the lowest occurrences of lexical complexity correlation – three instances for both. Of interest here are the negative correlations between C/T & VV1, and DC/C, DC/T and ADVV. As the ratio of dependent clauses per clause or t-unit increases, the range of verbs and adverbs used appears to decrease.

Table 3
Correlation between lexical complexity and subordination-based syntactic complexity measurements

LC Indices	**Amount of subordination**							
	C/T	Rs *(p)*	CT/T	Rs *(p)*	DC/C	Rs *(p)*	DC/T	Rs *(p)*
NDWZ-50			0.216 (0.035)		0.201 (0.045)		0.204 (0.042)	
VV1	-0.201 (0.045)							
ADVV					-0.217 (0.03)		-0.198 (0.048)	

The coordination indices (Table 4) reported no correlation with adjective or adverb measures. Particularly noteworthy are the high frequency of negative T/S correlations – each one is negative. This result suggests that the participants that produced more complex sentences through coordination were unable to vary their lexical choices or make use of more sophisticated vocabulary.

Table 4
Correlation between lexical complexity and coordination-based syntactic complexity measurements

LC Indices	**SC length of coordination indices**					
	CP/C	Rs *(p)*	CP/T	Rs *(p)*	T/S	Rs *(p)*
LD		-		-	-0.296 (0.003)	
LS1		-		-	-0.307 (0.002)	
LS2		-		-	-0.333 (0.001)	
CVV1	0.220 (0.028)			-		-
SVV1	0.221 (0.027)			-		-
NDW		-		-	-0.229 (0.022)	
NDWZ-50	0.211 (0.035)		0.198 (0.049)			-
NDW-ER50		-	0.256 (0.010)		-0.311 (0.002)	
TTR	0.228 (0.023)		0.205 (0.041)		-0.225 (0.025)	
CTTR	0.200 (0.047)			-	-0.288 (0.004)	
RTTR	0.200 (0.046)			-	-0.288 (0.004)	
LOGTTR	0.246 (0.014)		0.219 (0.028)		-0.271 (0.006)	
UBER	0.228 (0.023)			-	-0.280 (0.005)	
VV1	0.292 (0.003)		0.217 (0.030)			-

Table 5 displays the relationship between phrasal sophistication measures and lexical complexity. Firstly, note the absence of any correlation with the verb-phrase per t-unit index – it is unique in this regard, suggesting that it is not at all a reliable indicator of overall proficiency, at least in this particular context. Also of note is the prominence of CN/C with lexical measures. That CN/C correlates with all LC measures indicate that the use of complex nominals is a strong indicator of lexical complexity – writers that are able to construct sentences using complex nominals appear more likely to use a more varied and

sophisticated vocabulary. This is also possibly a result of task-effect – a writing task, such as the argumentative essay on which these measurements are based, encourages the use of more critical and analytical thinking, which in turn leads to the use of features that are more typical in academic uses of language (such as nominalisation). CN/C has a negative correlation with ADVV, suggesting that texts using more complex nominals also use a narrower range of adverbs.

Table 5
Correlation between lexical complexity and phrasal sophistication-based syntactic complexity measurements

LC Indices	**SC phrasal sophistication indices**					
	CN/C	Rs *(p)*	CN/T	Rs *(p)*	VP/T	Rs *(p)*
LD	0.295 (0.003)		0.287 (0.004)			-
LS1	0.325 (0.001)		-			-
LS2	0.339 (0.0006)		0.216 (0.031)			-
CVV1	0.314 (0.001)		-			-
SVV1	0.314 (0.001)		-			-
NDW	0.505 (<0.0001)		0.356 (0.0003)			-
NDWZ-50	0.292 (0.003)		0.342 (0.0005)			-
NDW-ER50	0.508 (<0.0001)		0.464 (<0.0001)			-
NDW-ES50	0.419 (<0.0001)		0.401 (<0.0001)			-
TTR	0.334 (0.0006)		0.281 (0.005)			-
MSTTR-50	0.426 (<0.0001)		0.371 (0.0001)			-
CTTR	0.534 (<0.0001)		0.405 (<0.0001)			-
RTTR	0.533 (<0.0001)		0.404 (<0.0001)			-
LOGTTR	0.403 (<0.0001)		0.328 (0.0009)			-
UBER	0.470 (<0.0001)		0.364 (0.0001)			-
LV	0.355 (0.0003)		0.223 (0.024)			-
NV	0.283 (0.004)		-			-
VV1	0.293 (0.003)		-			-
ADJV	0.396 (<0.0001)		-			-
ADVV	-0.241 (0.016)		-			-

Finally, table 6 displays the overall sentence complexity index against correlations with lexical complexity. As with T/S, the correlations are negative. This suggests that as the number of clauses per sentence increases, the level of lexical sophistication and verb variation decreases. Conversely, texts using more sophisticated and varied vocabulary use a more concise style of writing.

Table 6
Correlation between lexical complexity and overall sentence complexity measurements

LC indices	Overall sentence complexity C/S
LS1	-0.273 (0.006)
LS2	-0.258 (0.009)
VV1	-0.200 (0.049)

6 Conclusion

This study aimed to discover a unit of syntactic complexity that could also account for lexical complexity by investigating correlations between them. In total, 101 correlations between syntax and lexis were found. Of these, it was found that mean length of clause (MLC) and complex nominals per clause (CN/C) could potentially serve as indicators of overall linguistic complexity – at least within the specific context of this study.

However, the results raise questions regarding the utility and application of these measures. It is likely that their effectiveness is contingent on the proficiency level of the learners and the task they are completing. For example, learners need to be able to construct complex nominals effectively or be set a task that requires the use of such phrasal elaboration, for the measurements to have any use. Furthermore, the CN/C measure itself is not able to inform as to how the nominal is complex, or rather, how the complex nominal is formed.

Despite the aims of this research, the results appear to underline the limitations of relying on a single measure of complexity. While a single measure may offer simplicity and reduce the effects of measure-redundancy, it restricts the information available to the researcher and teacher. In light of this, it remains clear that a fine-grained, multifaceted approach to analysing complexity is more advantageous, in that a combination of lexical and syntactic complexity analyses are more likely to provide a deeper understanding of a learner's capabilities. This is congruent with the preliminary findings of my doctoral research, which so far indicate that complexity analysed through a set of lexicogrammatical developmental stages offers a more transparent, and didactically useful, interpretation of complexity and its relationship to proficiency.

References

Casal, E. J., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal*, *57*(9), 1318–1330. https://doi.org/10.1093/comjnl/bxt117

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing 28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In: A. Housen, F. Kuiken, & I. Vedder, (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA language learning & language teaching.* (pp. 1–20) John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.01hou

Hunt, K. W. (1965). *Grammatical structures written at three grade levels: NCTE Research Report No. 3.* National Council of Teachers of English.

Johnson, M. D., Acevedo, A., & Mercado, L. (2013). What vocabulary should we teach: Lexical frequency profiles and lexical diversity in second language writing. *Writing & Pedagogy*, *5*(1), 83–103. https://doi.org/10.1558/wap.v5i1.83

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, *7*(3), 700–715. https://doi.org/10.1002/tesj.238

Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27–51, https://doi.org/10.15858/engtea.69.4.201412.27

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Kovacevic, E. (2018). The relationship between lexical and syntactic complexity measures in a learner corpus. In S. Gudurić, & B. Radić-Bojanić, (Eds). *Jezici i culture u vremenu i prostoru* (Vol. 7, number 2; pp. 469–479). Filozofski fakultet Novi Sad.

Larsson, T., & Kaatari, H. (2020) Syntactic complexity across registers: Investigating (in) formality in second-language writing. *Journal of English for Academic Purposes*, *45*, 100850, 1–16. https://doi.org/10.1016/j.jeap.2020.100850

Li, H. (2015). Relationship between measures of syntactic complexity and judgments of EFL writing quality. In B. Ma, L. Cheng, H. He, L. Hale, & J. Zhang (Eds.), *Proceedings of the 2015 youth academic forum on linguistics, literature, translation and culture* (pp. 216–223). American Scholars Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

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_1.x

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515. https://doi.org/10.3758/s13428-012-0258-1

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly 15*(1), 29-43. https://doi.org/10.1080/15434303.2017.1405421

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*(2), 420–430. https://doi.org/10.1002/tesq.91

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing*, *33*, 3–17. https://doi.org/10.1016/j.jslw.2016.06.001

Williams, C. (2023). It's complicated: The relationship between lexis, syntax, and proficiency. *Discourse and Interaction*, *16*(2), 124–144. https://doi.org/10.5817/DI2023-2-124

Yang, W., Lu, X., & Cushing Weigle, S. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Appendix

Lexical complexity measures as used in the LCA (Lu, 2012)

Index	Label	Calculation	Explanation
Lexical density	LD	N_{lex}/N	Lexical words to the number of words
Lexical sophistication-I	LS1	$Nslex / N_{lex}$	Sophisticated lexical words to the total number of lexical words
Lexical sophistication-II	LS2	T_s/T	Sophisticated word types to the total number of word types
Verb sophistication-I	VS1	$Tsverb / Nverb$	Number of sophisticated verb types to the total number of verbs
Verb sophistication-II	VS2	$T2sverb / Nverb$	Variations (corrections) of VS1 measure
Corrected VS1	CVS1	$Tsverb / \sqrt{2}Nverb$	
Number of different words	NDW	T	Number of different words used in a language sample
NDW (first 50 words)	NDWZ-50	T in the first 50 words of sample	
NDW (expected random 50)	NDW-ER50	Mean T of 10 random 50-word samples	
NDW (expected sequence 50)	NDW-ES50	Mean T of 10 random 50-word sequences	
Type/Token ratio	TTR	T/N	Number of word types to the number of words in a text
Mean Segmental TTR (50)	MSTTR-50	*Mean segmental TTR – 50-word non-overlapping segments*	
Corrected TTR	CTTR	$T_{sverb} / \sqrt{2}N$	
Root TTR	RTTR	$T_{sverb} / \sqrt{N}$	
Bilogarithmic TTR	logTTR	$LogT / LogN$	
Uber Index	Uber	$Log^2 N / Log (N/T)$	
Verb variation-I	VV1	T_{verb} / N_{verb}	Variation of specific classes of words
Squared VV1	SVV1	T^2_{verb} / N_{verb}	
Corrected VV1	CVV1	$T_{verb} / \sqrt{2}N_{verb}$	
Lexical word variation	LV	T_{lex} / N_{lex}	
Verb variation-II	VV2	T_{verb} / N_{lex}	
Noun variation	NV	T_{noun} / N_{lex}	
Adjective variation	AdjV	T_{adj} / N_{lex}	
Adverb variation	AdvV	T_{adv} / N_{lex}	
Modifier variation	ModV	$(T_{adj} + T_{adv}) / N_{lex}$	