Corpus-Based Lexical Development of EFL Writing

Weilu Wang[1], Jijun Wang[1], and Manfu Duan[2(✉)]

[1] Foreign Languages College, Inner Mongolia University, Hohhot, People's Republic of China
[2] Division of International Cooperation and Exchange, Inner Mongolia University, Hohhot, People's Republic of China
duanmanfu@126.com

Abstract. Learner corpora, with their detailed information on learner language use, have been widely explored in second language acquisition and teaching. This paper is based on a self-built longitudinal EFL learner corpus to partly meet a long-desired goal of measuring and describing the general guiding feature and the dynamics of learner language, especially for beginners. The current study uses NLP tools to calculate the values for the variables needed for measuring lexical development: types, tokens, TTR, unit length indices, COCA frequency list coverage, and lexical sophistication indices. As the data are in abnormal distribution, independent-samples Kruskal-Wallis tests are employed to test the significance; further pairwise comparisons are to determine the difference between group pairs by year. The present study finds that conventional global variables are more applicable for learner language development for beginners, including the number of tokens and types, the number of letters per word and the number of words per sentence, bigram frequency, and bigram mutual information. At the same time, some of the novel indices do not make significant differences, such as TTR, MATTR, MTLD, MTLD-Ma-Wrap, COCA frequency list coverage, trigram frequency and trigram mutual information. The present study also notes that spelling mistakes hinder statistical accuracy in processing beginner language. The real difficulty of beginners lies in their lack of knowledge and practice of non-literary, suggestive or affective use of content words; correct use of topic-specific words, lexical bundles, and set collocations also pose great challenges. The findings provide new insights into EFL learner language and offer helpful pedagogical implications.

Keywords: Lexical development · EFL writing · Longitudinal learner corpus

1 Introduction

Lexical items form a vital part of any natural human language, and the mastery of verbal use dictates the linguistic performance of a language learner. An extensive vocabulary constitutes a language learner's ultimate objective (Yang and Coxhead 2020). As a result, lexical use and development have been a significant issue in language acquisition, teaching, and computational linguistics, providing implications for textbook and dictionary compilation.

Y. Yau and F. Hage Chehade (Eds.): DESD 2022, ASSEHR 691, pp. 53–65, 2023.
https://doi.org/10.2991/978-2-494069-37-4_9

A learner corpus puts together the data of actual learner language use, which provides detailed and accurate information on the specific features of learner language and learner difficulty. A close description of learner language can be achieved by collecting statistical data on learners' actual use of a language. At the same time, learner difficulty can be revealed by comparing EFL learner language with that of native speakers. Lexical study of a learner corpus of a specific group of learners facilitates our understanding of the peculiarities of learner language, which helps in the compilation of individualized learning material and in the understanding of common difficulties that various learners may meet in their lexical development.

Over the years, a great variety of corpora of learners from a wide range of cultural backgrounds have been built worldwide in countries like Belgium, Japan, Hungary, USA., UK., Poland, and China. The most well-known should be the Cambridge CLC, composed of over 1.6 million words collected from compositions by learners from 86 different cultural backgrounds. And 600,000 words in the CLC corpus have been annotated with errors annotations. Another one worth mentioning is the International Corpus of Learner English (ICLE), 2 million words of written language with abundant annotations such as POS, syntactic and error annotations. Meanwhile, many more efforts have been directed to small-scale self-built learner corpora for specific research purposes. Data drawn from such corpora have been widely used for describing and analyzing a string of aspects of the actual use of learner language.

As instructors of English, we generally believe that learning is seldom an intentional pursuit on the part of the learner, instead, it is the natural consequence of carrying out appropriate learning activities. And these activities should be centered on a clear understanding of the actual difficulties of the target learners. Therefore, a detailed description of the specific features and the developmental dynamics of learner language form the core of language teaching and learning. With the help of a learner corpus and the availability of natural language processing tools (NLP) and the accessibility of hypothesis testing programs, the ultimate prospect of describing, measuring, and analyzing learner language is made possible. But for beginners, the task seems even more enormous, as the biggest challenge in learning a language is the familiarity with language symbols and the mastery of the guiding rules; consequently, beginners are struggling in these aspects. With a clear picture of learner difficulty in mind, an instructor is equipped with a better means to prepare the learners for progress and pave their way to level up. The responsibility is on the academia to focus on specific traits of learner language, to detect their significant problems, and to discern and determine the factors that shape their writing quality. In light of this, the present study aims to provide a clearer picture of beginners' lexical use in EFL writing.

The peculiarities of learner language are first and foremost embodied in the use of specific individual words; therefore, the disjunctions in the use of such specific lexical items between EFL learners and native speakers arouse great interest from researchers. Xu (2016) explores the double object construction of the verb "give" and examines its use in EFL learners. The study finds that learners acquire the usage in a specific order: in terms of indirect object, pronouns come first and nominal phrases settle in last; for direct object, learners begin with short single-word constructions and later develop to longer ones of complex construction.

Another aspect that represents the distinctiveness of learner language lies in their use of lexical items of a particular type. Deshors (2015) taps into the use of gerund and infinitive used as verb complements in writing by EFL learners, and argues that in choosing a gerund or an infinitive, learners may concern more about the global grammatical context rather than focus on meaning only. Zhang and Mai (2017) examine the shaping factors on the acquisition of denominal verbs, and they assert that entrenchment and preemption significantly affect learner acceptance of denominal verbs. Babanoğlu (2018) compares ESL learners of Turkish and German on their distinguished character in the use of motion verbs in two learner corpora. The findings indicate that the use of motion verbs is significantly influenced by the first language: Turkish learners use more motion verbs of manner and demonstrate a greater variety in terms of vocabulary choice. Bartley & Tenorio (2016) explore the use of model verbs in a learner corpus; they assert that the certainty of stand-taking is closely related to gender, language proficiency and familiarity with the genre at hand. Previous studies on lexical items provide us with specific detailed information on the actual use of learner language; however, an overall holistic description of learner language and its development is also needed, which facilitates our general understanding of the developmental features and dynamics of learner language.

Research on lexical features of learner language mainly centers on lexical complexity, sophistication and variety. As Wesche & Paribakht (1996) and Wang (2014) suggest, vocabulary knowledge is generally represented in both breadth and depth. Breadth is mainly measured by using the number and frequency of words. At the same time, depth can be calculated by a large variety of indices, including word range, n-gram frequency, n-gram range, n-gram strength of association, contextual distinctiveness, semantic network and word neighbors (Kyle et al., 2015, 2018; Garner et al., 2019). Based on previous research, the present study takes indices like the number of words, frequency of words and the use of n-gram to describe lexical features of written language in an EFL learner corpus and the longitudinal developmental dynamics.

2 Method

2.1 Corpora

This study analyzes a self-built English learner corpus of 73,000 words, which consists of compositions by EFL beginners in a university in China. The language data for the corpus were collected during an English learning program of six semesters (three years). They were taken from the writing tasks in the end-of-semester exam for each semester. The data were then scanned, recognized, cleaned, discerned and checked, and texts which are too short or irrelevant to the topic suggested in the tests were written off.

2.2 Data Collection and Data Analysis

The data for this study is collected with natural language processing tools (NLP) designed by Christopher Kyle, an assistant professor of linguistics at the University of Oregon, and Scott Crossley, a professor of applied linguistics in the Department of Applied Linguistics and ESL, Georgia State University. Among them there are SiNLP, TAALED and

TAALES. SiNLP is a simple tool that allows users to analyze texts with an individualized dictionary. It also provides the user with general information on the text processed, including the number of words, the number of types, TTR, the number of letters per word, the number of paragraphs, the number of sentences, and the number of words per sentence for each text. For the dictionary to be compared, COCA vocabulary frequency list is used. Another one for the research purpose is TAALED, an analysis tool to calculate a wide variety of lexical diversity indices, which are calculated by using lemma forms such as all lemmas, content lemmas, or function lemmas. And there is TAALES, which is a tool that measures over 400 classic and new indices of lexical sophistication, and it includes indices related to a wide range of sub-constructs. In addition, it provides comprehensive index diagnostics, including text-level coverage output (the percent of words, bigrams, trigrams in a text covered by the index) and individual words, bigram, and trigram index coverage information.

3 Results

Based on previous research (Hashimoto, 2019; Nurmukhamedov, 2021; Crossley et al., 2011; Kim et al., 2018), this paper chooses indices like the number of tokens, the number of types, TTR, the number of letters per word, the number of words per sentence, MATTR, MTLD, MTLD-Ma-Wrap and coverage percentage of COCA frequency list (high-frequency, mid-frequency and low-frequency) as indices for describing the lexical diversity and complexity of beginner language development.

3.1 Types, Tokens and TTR

This paper describes the overall characteristics of learner language by indices like the number of tokens, the number of types and TTR. Table 1 shows the descriptive statistics of the number of types and tokens and TTR values. As is indicated in the table, from the first year to the third, the mean values of the three indices are generally on the rise, which suggests that the learners are using more words and more complex forms longitudinally. Moreover, it can be noticed that the standard deviations over the semester are growing as well, which denotes the fact that individual differences are also widening. And later Kruskal-Wallis tests for several independent samples, as shown in the same table, indicates significant differences among the groups by year (sig = 0.00 for tokens and types, sig = 0.008 for TTR).

Further pairwise comparisons for the three indices show that for the number of tokens and types, there are significant differences among the data collected in the three years. It is a sure sign that EFL learners write longer compositions and use more words in their writing. Therefore, the number of tokens and types can be reliable indicators for learner language development. Regarding the type-token ratio (TTR), there are significant differences between the data in the first year and the second and between the first and third. In contrast, there are significant differences in 1st Year- 2nd Year and 1st Year- 3rd Year pairs; in contrast, there is no significant difference in the 2nd Year- 3rd Year pair. It shows that compared with the preliminary level of the first year, the learners are making progress in the second year and the third. But compared with the second year, learners in the third year are not doing significantly better in lexical diversity.

Table 1. Descriptive statistics, results of Kruskal-Wallis tests and pairwise comparisons for tokens, types and TTR (by year)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year – 2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
Tokens	1st Year	136	139.955	27.529	.000*	.000*	.025*	.000*
	2nd Year	165	171.321	52.011				
	3rd Year	213	181.098	57.040				
	Total	514	164.125	45.527				
Types	1st Year	136	83.595	15.040	.000*	.000*	.003*	.000*
	2nd Year	165	96.957	24.105				
	3rd Year	213	107.694	30.954				
	Total	514	96.082	23.366				
TTR	1st Year	136	.602	.066	.008*	.010*	1.000	.036*
	2nd Year	165	.577	.063				
	3rd Year	213	.570	.122				
	Total	514	.580	.094				

3.2 Unit Length

Table 2 is the descriptive statistics for two other commonly used global indices for lexical complexity: the number of letters per word and the number of words per sentence. As suggested in the table, the mean values for the two indices get more prominent from the first year to the third year (3.56, 3.82 and 4.0 for the number of letters per word; 11.69, 13.04 and 16.41 for the number of words per sentence).

Kruskal-Wallis tests for the two indices indicate significant differences among the year groups (sig. = 0.000). The results of further pairwise comparisons show that there are significant differences for both indices (sig. = 0.000 for the number of letters per word; sig. = 0.000 for the number of words per sentence for pairs of 1st Year- 3rd Year and 2nd Year- 3rd Year, sig. = 0.021 for the 1st Year- 2nd Year pair). The results certify that indices of unit length are reliable indicators for learner language development.

Table 2. Descriptive statistics, results of the Kruskal-Wallis tests and pairwise comparisons for Letters per word(L/W) and Words per sentence (WPS)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test	Pairwise Comparisons		
						1st Year - 2nd Year	2nd Year - 3rd Year	1st Year - 3rd Year
L/W	1st Year	136	3.5601	.27153	Sig. = .000*	Sig. = .000*	Sig. = .000*	Sig. = .000*
	2nd Year	165	3.8210	.37344				
	3rd Year	213	4.0287	.36759				
	Total	514	3.8380	.39432				
WPS	1st Year	136	11.6863	5.06783	Sig. = .000*	Sig. = .021*	Sig. = .000*	Sig. = .000*
	2nd Year	165	13.0423	5.66668				
	3rd Year	213	16.4058	6.51964				
	Total	514	14.0773	6.22103				

3.3 COCA Frequency List Coverage

The present study uses a user dictionary tailored from COCA lemma frequency list. And the lemmas are grouped into three categories: high-frequency, mid-frequency and low-frequency. The first two thousand most frequently used lemmas are put into the group of high-frequency, the following three thousand on the frequency list are grouped as mid-frequency, while the 5001st to 10000th on the list are classified as low-frequency. Table 3 shows the descriptive statistics for COCA frequency list coverage. It can be noticed that a large part of learner language falls into the high-frequency group (mean > 0.6), while only a minor proportion of words is within the low-frequency (mean < 0.05).

The results of nonparametric Kruskal-Wallis tests for the three pairs are shown in the same table, indicating significant differences. And subsequent pairwise comparisons provide further information: for the groups of high-frequency and mid-frequency, significant differences can only be observed for the 1st Year - 3rd Year pair (sig. = 0.25 and 0.12, respectively); while for the low-frequency group, there are significant differences in the 1st Year – 2nd Year and the 2nd Year – 3rd Year pairs, but somewhat unpredictably no significant difference is noticed in the 1st Year – 3rd Year pair. It seems that COCA frequency list coverage is not so consistent in predicting the development of learner language.

Table 3. Descriptive statistics, results of Kruskal-Wallis tests and pairwise comparisons for COCA Frequency List Coverage

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test	Pairwise Comparisons		
						1st Year – 2nd Year	2nd Year – 3rd Year	1st Year – 3rd Year
High-Freq	1st Year	136	.6392	.06551	Sig. = .027*	Sig. = .829	Sig. = .350	Sig. = .025*
	2nd Year	165	.6268	.07022				
	3rd Year	213	.6096	.09504				
	Total	514	.6230	.08116				
Mid-Freq	1st Year	136	.1085	.03233	Sig. = .000*	Sig. = .215	Sig. = .892	Sig. = .012*
	2nd Year	165	.1017	.03115				
	3rd Year	213	.0988	.03360				
	Total	514	.1023	.03267				
Low-Freq	1st Year	136	.0800	.02357	Sig. = .0115*	Sig. = .000*	Sig. = .000*	Sig. = .068
	2nd Year	165	.0532	.02447				
	3rd Year	213	.0796	.04445				
	Total	514	.0712	.03616				

3.4 Lexical Sophistication

Modeled on previous research (Zenker & Kyle, 2021; Covington & McFall, 2010; McCarthy, 2005; McCarthy & Jarvis, 2010), the current study takes lexical density, MATTR50 (moving average type-token ratio), MTLD (the average number of tokens needed for a given TTR value: 0.720) and MTLD-Ma-Wrap (moving average version of MTLD) as indices for measuring lexical sophistication of learner language development. The descriptive statistics for the four variables are shown below in Table 4. In terms of mean, most indices rise from the first year to the third year.

Later Independent-samples Kruskal-Wallis tests shown in the same table demonstrate that there are significant differences existing over the years for the first three measurements, namely lexical diversity (types), lexical diversity (tokens), and MATTR50 (sig. = 0.00 for lexical diversity, sig. = 0.006 for MATTR50), while no significant difference exists for MTLD and MTLD-Ma-Wrap. Further pairwise comparisons indicate that there are significant differences for pairs 1st Year- 2nd Year and 1st Year- 3rd Year for lexical diversity (sig. = 0.000 and 0.003 for types and sig. = 0.002 and 0.000 for tokens), while significance can only be noticed for the 2nd Year - 3rd Year pair for MATTR50. The

Table 4. Descriptive statistics, results of Kruskal-Wallis tests and pairwise comparisons for Lexical Sophistication

Variables Year		N	Mean	Std. Deviation	Kruskal-Wallis test	Pairwise Comparisons		
						1st Year - 2nd Year	2nd Year - 3rd Year	1st Year – 3rd Year
lexical_density_types	1st Year	136	0.558	0.049	Sig. = .000*	Sig. = .000*	Sig. = .072	Sig. = .003*
	2nd Year	165	0.588	0.056				
	3rd Year	213	0.575	0.071				
	Total	514	0.575	0.062				
lexical_density_tokens	1st Year	136	0.427	0.049	Sig. = .000*	Sig. = .002*	Sig. = .315	Sig. = .000*
	2nd Year	165	0.446	0.053				
	3rd Year	213	0.458	0.075				
	Total	514	0.446	0.063				
mattr 50_aw	1st Year	136	0.712	0.058	Sig. = .006*	Sig. = 1.000	Sig. = .007*	Sig. = .070
	2nd Year	165	0.712	0.050				
	3rd Year	213	0.725	0.059				
	Total	514	0.717	0.056				
mtld_original_aw	1st Year	136	48.558	16.248	Sig. = .137	----	----	----
	2nd Year	165	47.593	13.915				
	3rd Year	213	50.322	16.275				
	Total	514	48.942	15.542				
mtld_ma_wrap_aw	1st Year	136	47.101	15.632	Sig. = .155	----	----	----
	2nd Year	165	46.484	13.721				
	3rd Year	213	49.002	16.423				
	Total	514	47.654	15.371				

results suggest that the five variables tested here are not applicable for measuring learner language development.

3.5 Bigram and Trigram Complexity

As suggested by Davies (2009), the present research takes COCA academic bigram lemma frequency, range, mutual information, and COCA academic trigram lemma frequency, range and mutual information as variables for learner language development.

Table 5 shows the descriptive statistics for bigram and trigram complexity of learner language. As is indicated in the same table, the mean values for bigram frequency are all larger than 180 and are on the rise over the years, while the means for trigram frequency are over 11 and are also growing over the years. There is no apparent linear growth noticed in terms of mean values for all the four other variables.

As shown in the table, the results of the independent-samples Kruskal-Wallis Test indicate that there are significant differences for all the six variables (sig. = 0.000). And further pairwise comparisons (shown in the same table) indicate that for variables like bigram frequency, bigram mutual information, and trigram range, significant differences are noticeable over the year groups (sig. < 0.05). But for the remaining three, some of the pairs do not pose significant differences (sig. > 0.05). And it can be drawn that variables for bigram are more applicable than those for trigram.

Table 5. Descriptive statistics, results of Kruskal-Wallis tests and pairwise comparisons for N-gram complexity

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test	Pairwise Comparisons		
						1^{st} Year – 2^{nd} Year	2^{nd} Year – 3^{rd} Year	1^{st} Year – 3^{rd} Year
COCA_Academic_ Bigram_Lemma_ Frequency	1st Year	136	188.22	124.021	Sig. = .000*	Sig. = .003*	Sig. = .045*	Sig. = .000*
	2nd Year	165	221.315	114.192				
	3rd Year	213	258.808	183.181				
	Total	514	227.236	150.284				
COCA_Academic_ Bigram_Lemma_ Range	1st Year	136	0.165	0.118	Sig. = .000*	Sig. = .000*	Sig. = .133	Sig. = .000*
	2nd Year	165	0.210	0.241				
	3rd Year	213	0.178	0.048				
	Total	514	0.185	0.155				
COCA_lemma_ academic_bi_MI	1st Year	136	1.297	0.219	Sig. = .000*	Sig. = .013*	Sig. = .000*	Sig. = .000*
	2nd Year	165	1.206	0.222				
	3rd Year	213	1.044	0.567				
	Total	514	1.166	0.410				
COCA_lemma_ Academic_Trigram _Frequency	1st Year	136	11.113	8.311	Sig. = .000*	Sig. = .100	Sig. = .010*	Sig. = .000*
	2nd Year	165	12.141	7.152				
	3rd Year	213	15.371	9.921				
	Total	514	13.147	8.831				

(*continued*)

Table 5. (*continued*)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test	Pairwise Comparisons		
						1st Year – 2nd Year	2nd Year – 3rd Year	1st Year – 3rd Year
COCA_lemma_ Academic_Trigram_Range	1st Year	136	0.028	0.016	Sig. = .000*	Sig. = .037*	Sig. = .002*	Sig. = .000*
	2nd Year	165	0.032	0.015				
	3rd Year	213	0.039	0.022				
	Total	514	0.034	0.019				
COCA_lemma_ academic_tri_MI	1st Year	136	2.681	0.531	Sig. = .000*	Sig. = .001*	Sig. = .682	Sig. = .000*
	2nd Year	165	2.455	0.428				
	3rd Year	213	2.333	0.724				
	Total	514	2.468	0.603				

4 Discussion and Conclusion

The present study examines a range of variables for lexical features of learner language and explores their availability for longitudinal beginner language development over three years. As the data are in abnormal distribution, a series of independent-samples Kruskal-Wallis tests are employed to determine the significance, the results of which indicate that some widely practiced global indicators are more applicable for predicting language development for beginners; these indicators include the number of tokens and types, the number of letters per word and the number of words per sentence, bigram frequency, and bigram mutual information. But some of the new and novel indices are crossed out for their unavailability to measure beginner language development, such as TTR, COCA frequency list coverage, lexical diversity, and the use of trigrams.

The findings of this paper provide partial support for previous research on the relationship between features of lexical use and L2 writing development (Monteiro & Crossley, 2020; Mckee et al., 2000), while it casts doubts on the effectiveness and predictability of some newly-proposed indices in some other research (Engber, 1995; Kyle & Crossely, 2015), including TTR, COCA frequency list coverage, lexical diversity, and MATTR. The discrepancy can be explained by the fact that previous research is based on corpora collected from writing by intermediate or advanced learners, while the present study is focused on beginners' writing. As suggested by Pourdana et al. (2011), accuracy, fluency and complexity of learner language are influenced by task types. Due to a noticeable limit on proficiency, the appropriate writing tasks for beginners could be note writing or letter writing, as is the case in this study, while for learners of a higher proficiency level, expository, argumentative, and narrative assignments are acceptable. Therefore, the complexity and difficulty of the writing tasks and the different proficiency level of the learners are factors bringing in the disjunctions.

Still, there are some striking peculiarities of the lexical use of beginner language. On the frequency list of beginners, the most frequently used content words are words of less

than five letters, usually confined to their literal senses and fundamental notions. And more surprisingly, most of the complex or low-frequency words are wrong spellings. Even those simple words (high-frequency words), verbs mainly, are usually limited to a smaller number of collocations and constructions compared with advanced learners. Furthermore, there stands no chance to observe phrasal verbs and set phrases or figures of speech in beginner writing. These aspects of language use are the genuine difficulties for beginners, and these aspects form the significant threshold for their leveling up to intermediate. The findings call our attention to a focus on these aspects of lexical use in the teaching process, the writing of textbooks, and the design of the curriculum. A beginner can begin with the "Hello" and "How-are-you" kind of language learning, but they have to be going on to outgrow that stage over the years, which is especially true for learners at the college level and adults. Lexical bundles, idiomatic usage, and set collocations are an indispensable part of language skill-building, and due attention is supposed to be paid to these aspects.

The present research discusses the availability of a range of variables for measuring the longitudinal development of EFL beginner writing based on a corpus collected from an English program for college students. The study provides new insights into beginner language features regarding lexical use, namely diversity and complexity. It talks about the overall development of learner language, and future studies can further explore the development of students of different learning abilities and proficiency levels. And as shown by the descriptive statistics, the values for standard deviation over the years for some variables are getting larger, indicating a widening gap among learners. Questions like how come the widening gap and what the gap is really like deserve our attention. This study analyzes data collected over three years only; a more extended observation period is needed to check the result for a longitudinal study. The data for the corpus of this study are written language, and other forms of learner language production call for more research, such as oral English or translation English. Furthermore, learner language may be constantly influenced by language input, including instructors' classroom instruction, feedback, peers' responses, and learning materials online or in textbooks, on which further studies are greatly needed.

Acknowledgments. The present research is funded by the National Social Science Fund of China (17BYY042, A Typological Study of the Function-order Interactions of the Modifiers of English and Chinese) and the Innovative Practice Base for Developmental Integration of Information Technology with Foreign Languages Teaching and Research.

References

1. BABANOĞLU M P, Motion Verbs in Learner Corpora, Gaziantep University Journal of Social Sciences, 2018, 17(1), pp. 221–228. DOI: https://doi.org/10.21547/jss.372590
2. Bartley L, Hidalgo-Tenorio E, "Well, I think that my argument is…," or modality in a learner corpus of English, Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics, 2016, 29(1), pp. 1-29. DOI: https://doi.org/10.1075/resla.29.1.01bar
3. Covington M A, McFall J D, Cutting the Gordian knot: The moving-average type–token ratio (MATTR), Journal of quantitative linguistics, 2010, 17(2), pp. 94-100. DOI: https://doi.org/10.1080/09296171003643098

4. Kyle K, Crossley S, Berger C, The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0, Behavior research methods, 2018, 50(3), pp. 1030-1046. DOI: https://doi.org/10.3758/s13428-017-0924-4
5. Crossley S A, Salsbury T, McNamara D S, et al, Predicting lexical proficiency in language learner texts using computational indices, Language Testing, 2011, 28(4), pp. 561-580. DOI: https://doi.org/10.1177/0265532210378031
6. Davies M, The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights, International journal of corpus linguistics, 2009, 14(2), pp. 159-190. DOI: DOI: https://doi.org/10.1075/ijcl.14.2.02dav
7. Deshors S C, A multifactorial approach to gerundial and to-infinitival verb-complementation patterns in native and non-native English, English Text Construction, 2015, 8(2), pp. 207-235. DOI: DOI: https://doi.org/10.1075/etc.8.2.04des
8. Garner J, Crossley S, Kyle K, N-gram measures and L2 writing proficiency, System, 2019, 80, pp. 176-187. DOI: https://doi.org/10.1016/j.system. 2018.12.001
9. Hashimoto B J, Egbert J, More than frequency? Exploring predictors of word difficulty for second language learners, Language Learning, 2019, 69(4), pp. 839–872. https://doi.org/10.1111/lang. 12353
10. Kim M, Crossley S A, Kyle K, Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality, The Modern Language Journal, 2018, 102(1), pp. 120-141. DOI: https://doi.org/10.1111/modl.12447
11. Kyle K, Crossley S A, Automatically assessing lexical sophistication: Indices, tools, findings, and application, Tesol Quarterly, 2015, 49(4), pp. 757-786. DOI: https://doi.org/10.1002/tesq.194
12. Kyle K, Crossley S, Berger C, The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0, Behavior research methods, 2018, 50(3), pp. 1030-1046. DOI: https://doi.org/10.3758/s13428-017-0924-4
13. Laufer B, Waldman T, Verb-noun collocations in second language writing: A corpus analysis of learners' English, Language learning, 2011, 61(2), pp. 647-672. DOI: https://doi.org/10.1111/j.1467-9922.2010.00621.x
14. McCarthy P M, Jarvis S, MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, Behavior research methods, 2010, 42(2), pp. 381-392. DOI: https://doi.org/10.3758/BRM.42.2.381
15. McCarthy P M, An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD), The University of Memphis, 2005.
16. Nurmukhamedov U, Sharakhimov S, Corpus-Based Vocabulary Analysis of English Podcasts, RELC Journal, 2021. https://doi.org/10.1177/0033688220979315.
17. Pourdana N, Karimi Behbahani M, Safdari M, The impact of task types on aspects of Iranian EFL learners' writing performance: Accuracy, fluency, and complexity, Proceedings of International Conference on Humanities, Society and Culture. 2011.
18. Wang Z. A Correlation Analysis on the Depth and Breadth of ESL Learners' Vocabulary Knowledge and Their Overall Linguistic Competence[J]. Theory & Practice in Language Studies, 2014, 4(12).
19. Wesche M, Paribakht T S, Assessing second language vocabulary knowledge: Depth versus breadth, Canadian Modern Language Review, 1996, 53(1), pp. 13–40. DOI: https://doi.org/10.3138/cmlr.53.1.13
20. Xiao L, Assessing the roles of breadth and depth of vocabulary knowledge in second language proficiency, Foreign Language Teaching and Research, 2007, 39(5): 352-359.

21. Xu Q. Item-based foreign language learning of give ditransitive constructions: Evidence from corpus research[J]. System, 2016, p. 63 65–76. DOI: https://doi.org/10.1016/j.system.2016.08. 008
22. Yang L, Coxhead A, A corpus-based study of vocabulary in the new concept English textbook series, RELC Journal, 2020. DOI: https://doi.org/10.1177/0033688220964162
23. Zenker F, Kyle K, Investigating minimum text lengths for lexical diversity indices, Assessing Writing, 2021. DOI: https://doi.org/10.1016/j.asw.2020.100505/
24. Zhang X, Mai C, Effects of entrenchment and preemption in second language learners' acceptance of English denominal verbs, Applied psycholinguistics, 2018, 39(2), pp.413-436. DOI: https://doi.org/10.1017/S0142716417000 406