Pak. j. life soc. Sci. (2025), 23(1): 2505-2515 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences

www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2025-23.1.00196

RESEARCH ARTICLE

Building the CELTEC Corpus: Assessing Lexical Complexity in Chilean Higher Education L2 English Learners

Pamela Saavedra Jeldres[1*], Lucía Ubilla Rosales[2], Belén Muñoz Muñoz[3]

[1]Universidad Católica de Temuco, Chile, Departamento de Educación e Innovación, Edificio Agustina Hidalgo, Temuco Chile

[2]Universidad de la Frontera, Chile

[3]Universidad del Bío-Bío, Chile

ARTICLE INFO

Received: Nov 18, 2024

Accepted: Jan 10, 2025

Keywords

Computational linguistics

Learner corpus research

Lexical complexity

L2 writing

Pre-service language teachers

***Corresponding Author:**

psaavedra@uct.cl

ABSTRACT

This article introduces the development, construction, and potential applications of a learner corpus—Chilean English Language Teacher Education Corpus (CELTEC)—comprising 404 texts written by English as a foreign language (EFL) pre-service teachers enrolled at nine universities in Chile. The study outlines the methodology for creating this pseudo-longitudinal corpus, facilitating replication. It includes three cohorts representing years 3, 4, and 5 of a five-year undergraduate programme. Additionally, the study examines the lexical complexity of the corpus texts, focusing on constructs such as lexical density, diversity, and sophistication. Data were collected using the corpus query language (CQL) in Sketch Engine and analysed with freely available tools, TAALES 2.2 and TAALED 1.4, to calculate indices of lexical complexity within a multidimensional framework. The results reveal a slight developmental trend in lexical complexity across the CELTEC cohorts. Lexical density is moderate, averaging between 40-49%, yet increases incrementally with each academic year. Lexical diversity also shows improvement across cohorts; however, this growth does not consistently correspond to higher lexical sophistication in the texts. These findings have implications for English for General Academic Purposes (EGAP) pedagogy, both within the study's context and in broader educational settings. Specifically, they underscore the critical need for more explicit instruction in advanced academic vocabulary to address learners' specific needs effectively.

BACKGROUND

Technological advancements in computational tools have revolutionised corpus linguistics, facilitating the analysis of extensive datasets – an otherwise impractical task when performed manually. These developments are particularly impactful in the areas of vocabulary acquisition and instruction, as vocabulary plays a crucial role in language learning. As Wilkins (1972, p. 111) noted that "while without grammar very little can be conveyed, without vocabulary nothing can be conveyed". Thus, assessing learners' lexical knowledge and vocabulary size is vital for educators to design effective lessons and select appropriate materials that align with their students' lexical needs.

In the context of English for Academic Purposes (EAP), the development of advanced writing skills is essential for success in academia within today's globalised world. This is especially pertinent for students in Initial English Language Teacher Education (IELTE) programmes, known as *Pedagogía*

en inglés in Chile. These future teachers are tasked not only with enhancing their own writing proficiency but also with acquiring the necessary strategies to foster their future students' writing development. Research underscores that "the major determinant of the vocabulary used in written production is the vocabulary size of the writer, especially if the writer is a second language learner with a relatively small vocabulary when compared to native speakers" (Laufer & Nation, 1995, p.301).

Despite its significance, writing is among the most challenging skills to develop, requiring time to strengthen both syntactic and lexical complexity while mastering diverse genres and text types. This challenge has attracted considerable attention in applied linguistics over the past decades, with learner corpus research (LCR) emerging as a pivotal methodology for investigating second language (L2) writing development, teaching, and learning, among other fields (Paquot, 2018).

LCR can be conducted using sophisticated yet user-friendly linguistic software, often freely available online, to analyse learners' writing, such analyses uncover developmental patterns, providing valuable insights that inform pedagogical decisions. They offer a deeper understanding of learners' lexical complexity, revealing both strengths and areas for improvement, and enabling comparisons with established benchmarks, such as native or L2 writing.

One promising application of LCR is the creation of localised learner corpora, which can directly guide the planning of L2 writing instruction. As Guilquin (2015) observes local corpora "invite teachers and students alike into the field of learner corpus research by making them both providers and beneficiaries, thus resulting in learner corpora being directly useful to those for whom, ultimately, they have been compiled" (p. 29).

Over the last three decades, LCR has significantly advanced the understanding of learner language by providing detailed descriptions of usage across various features (Paquot, 2018), examining developmental trajectories (Biber, 2011, 2020; Kyle, 2018; Staples & Reppen, 2016), and offering robust statistical measures to analyse data (Biber, 2011, 2016; Kyle & Crossley, 2016, 2018, 2020). There is now general consensus on the key methodological issues in learner corpus compilation such as the absolute necessity of including metadata related both to the learners and to the tasks (Granger, 2015).

A core focus of LCR is linguistic complexity, underpinned by the assumption that more proficient learners produce more linguistically complex language (Gray et al., 2019). Ellis and Barkhuizen (2015) define complexity as "the use of more challenging and difficult language … complexity is the extent to which learners produce elaborated language" (2005, p. 139). Within this framework, lexical complexity – comprising density, diversity, and sophistication – has emerged as a key construct for assessing L2 proficiency (Laufer & Nation, 1995; Read and Read, 2000). Wolfe-Quintero et al. (1998) state that second language writing is characterised by lexical complexity or lexical richness expressed through the sophistication and range of an L2 learners' productive vocabulary.

Therefore, this study analyses three constructs of lexical complexity lexical: density, diversity and sophistication. Lexical density measures the proportion of content words in a text, lexical diversity examines the range of unique words used, and lexical sophistication assesses the use of infrequent, advanced vocabulary. These indices provide valuable metrics for gauging learners' progress and informing instructional practices (Clavel-Arrotitia & Pennock-Speck, 2021, p.232).

As stated earlier, lexical density provides a measure of the proportion of lexical items (i.e., nouns, verbs, adjectives) compared to function words (prepositions, articles, determiners) in a text. Texts with higher lexical density are typically more complex and harder to understand than those with lower density. For instance, in a study conducted by Zheng (2015), beginner Chinese learners of English exhibited approximately 60% lexical diversity in their texts, a result similar to those reported

by Schnur & Rubio, (2021) with a 60.88% in low proficient students (A1), and Zhang (2022), who found values around 57%.

On the other hand, "lexical diversity refers to the range of different words used in a text, with a greater range indicating a higher diversity" (McCarthy & Jarvis, 2010, p.381). A greater diversity suggests a broader vocabulary and is often measured using indices such as the type-to-token ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and Vocabulary Diversity (Vocd) (McCarthy & Jarvis, 2007, 2010). Research shows that lexical diversity positively correlates with writing performance (Crossley & McNamara, 2012). For instance, Zhang (2022) used the MTLD to measure lexical diversity in Chinese college-level essays, with a mean value of 71.51, which is considered relatively high - the closer to 100 the more lexically diverse the text is.

The construct of lexical sophistication reflects the depth and breadth of lexical knowledge available to speakers, readers, and writers (Meara, 2005 in Kyle & Crossley, 2015). Sophistication encompasses the use of less frequent lexis; generally, "words that appear in fewer texts are considered more sophisticated" (Kyle et al., 2018, p.1031). Academic language often employs specialised vocabulary that is less frequent in everyday language. Therefore, the presence of low-frequency, advanced words in learners' writing is a key indicator of higher proficiency (Crossley et al., 2014; Kyle & Crossley, 2015). However, L2 learners tend to rely more on high-frequency, conversational words (Laufer & Nation, 1995).

To the best of our knowledge, corpus research in Chile is in its infancy; thus, this investigation represents the first national learner corpus of English as a foreign language. By collecting, building and analysing lexical resources of local learners, this study offers a significant contribution to language teacher educators and pre-service teachers of IELTE programmes. Hence, the present study addresses the following research question:

What developmental trends regarding lexical complexity – encompassing lexical density, diversity, and sophistication – can be observed in learner writing from the CELTEC learner corpus across cohorts?

Materials and Methods

The main task of the methodology section was to build the learner corpus. This section will include the corpus design, data collection and compilation, and data analysis.

Corpus design

The design of the CELTEC corpus was based on three main general principles: a) characteristics of the corpus, b) variables related to the learners, and c) variables related to the task. The design is summarised in Table 1. As Granger (2004, p.126) notes, "one must admit that ... there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control". Therefore, this corpus incorporates different variables to contribute to its breadth, which might be used depending on the research objectives.

Table 1: Corpus design

Characteristics of the corpus	**Variables related to the learners**	**Variables related to the task**
1. **Written** mode	1. **Learning context** – Future teachers of English	1. **Genre** – self-reported text types
2. **Monolingual** – English as a foreign language	2. **Learner's proficiency level in L2**	2. **Number of words**

3. **Time** – Pseudo-longitudinal	3. **Learner's cohort** (third, fourth, fifth year)	3. **Type of task** – exam, class writing, free writing
4. **Breadth** – Specialised	4. **Learner's university** (State – subsidised – private)	
5. **Annotation** – POS, manual		

As summarised in Table 1, this corpus consists of the written work of students enrolled in Pedagog*ía en inglés* programmes*.* The texts were produced as part of various courses, including modules in language, linguistics, and didactics. The corpus is monolingual, comprising exclusively texts written in EFL.

Following Ellis and Barkhuizen (2005), the current study adopts a pseudo-longitudinal design since it provides an alternative in the absence of longitudinal learner data. Specifically, Ellis and Barkhuizen (2005) explain that: "samples of learner language are collected from groups of learners of different proficiency levels at a single point in time. A longitudinal picture can then be constructed by comparing the devices used by the different groups ranked according to their proficiency" (p. 97).

The corpus annotation was carried out using part-of-speech (POS) tagging via Sketch Engine. Furthermore, five variables were annotated manually for each of the 404 texts to facilitate detailed analyses: year of study, proficiency level, type of university, genre, task type.

Learner Context and Participants

The participants in this study are 139 undergraduate students enrolled in IELTE programmes in Chile. These programmes use English as a medium of instruction (EMI), in a predominantly Spanish-speaking country, reflecting a high level of motivation among these learners compared to other learner groups. Over the course of their five-year undergraduate studies, students undertake courses in linguistics and English Language Teaching (ELT) methodology to prepare for their future roles as secondary school teachers of EFL.

The **CELTEC corpus** encompasses 2.3 written contributions per participant who had diverse proficiency levels in the target language:

- **Year 3 (Level B1)**: 68 participants
- **Year 4 (Level B2)**: 279 participants
- **Year 5 (Level C1)**: 53 participants

These proficiency levels are aligned with the Common European Framework of Reference for Languages (CEFR). Notably, Level B2 is the most represented in the corpus, while Levels B1 and C1 comprise smaller sample subsets.

Data collection and compilation

An email was sent to the heads of the departments from the 34 universities in Chile that offer the IELTE programme. A link to a Google form was sent, and students were able to upload an unlimited number of texts together with their personal data (name, cohort, and university). Once the data collection process was accomplished, all texts underwent a cleaning procedure to delete any images, figures, charts, and personal information. The personal information was replaced by placeholder items, such as these: *<institution>, <city name>*. Moreover, comments and feedback from tutors had to be eliminated and track changes rejected very carefully so as to keep the original student text intact. Texts sent in PDF format needed to be converted into Word, and then the same cleaning process was applied accordingly to get plain texts. On the other hand, scanned texts, photographs of typed texts, or photographs of handwritten texts were typed in Word format and then saved as text

plain (McEnery & Hardie, 2012). Finally, a plagiarism check was conducted using a software called Turnitin. Texts with 25% similarity match or higher were manually checked. If the similarity was due to in-text citations, the text was included in the corpus. If not, it was excluded from the corpus. Encouragingly, only 23 compositions out of 427, accounting for 5.38% of the texts, were excluded due to heavy copying.

The annotation process was conducted on a text-by-text basis, covering all 404 texts in the dataset. This detailed annotation enabled the creation of sub-corpora within the CELTEC, facilitating more in-depth data analysis. Each text was assigned a unique code encapsulating five key variables. For instance, the code **celtec_analysis001_L087_assig_y4_b2_StU3** represents the following information: the text type is an analysis (text number 1), written by learner 87, as part of an assignment in year 4, with an expected proficiency level of B2, and originated from a state university designated as "3".

Table 2 summarises the main features of the CELTEC corpus, including its three subcorpora corresponding to cohorts in years 3, 4, and 5. The table highlights the number of words and texts per cohort, along with the most represented genres within each group, being the most frequent EFL genres include the essay, reports and reviews.

Tabla 2: Features from CELTEC

Criteria	**Year 3 (B1)**	**Year 4 (B2)**	**Year 5 (C1)**
Number of words	28,335	170,116	42,007
Mean Text length	463.2	574.5	820.6
Number of texts	56	291	57
Essays	15,259 (6.3%)	77,238 (31.8%)	27,028 (11.1%)
Reports	3,152 (1.3%)	17,005 (7%)	4,231 (1.7%)
Reviews	2,395 (1%)	7,891 (3.2%)	0
Proposals	1,130 (0.5%)	4,091 (1.7%)	1,743 (1.7%)
Letters	167 (0.1%)	3,811 (1.6%)	1,727 (0.7%)
Instructions	3,640 (1.5%)	0	0

Table 3 provides a detailed overview of the word ranges observed in CELTEC across cohorts, with most texts being placed in the bands between 201-400 words, followed by the band of 401-700 words.

Table 3: Word-bands across cohorts

	Number of texts		
Range of words	**Year 3**	**Year 4**	**Year 5**
0-200	13 (23.2%)	20 (6.8%)	0
201-400	25 (44.6%)	123 (42.2%)	15 (26.3%)
401-700	10 (17.8%)	74 (25.4%)	13 (22.8%)
701-1000	4 (7.1%)	32 (10.9%)	11 (19.2%)
1000-2000	2 (3.5%)	35 (12%)	12 (21)
2000-3000	2 (3.5%)	6 (2%)	2 (3.5%)
3000 or higher	0	1 (0.3)	3 (5.2%)
Total number of texts	**56**	**291**	**57**

Data analysis

Data analyses were carried out using a multidimensional approach to measure lexical density, diversity and sophistication. Lexical density was calculated using Halliday's (1985) formula, which

divides the number of content (lexical) tokens by the total number of tokens, multiplied by 100. This information was provided by the freely available software TAALED 1.4 (Kyle & Crossley, 2015) and repeated for Year 3, Year 4 and Year 5 cohorts. Lexical diversity used indices from TAALES 2.2 (Kyle & Crossley, 2018). The indices included the basic number of tokens, types (of words), content types, function types, lexical density types, the simple type-token ratio for all words (TTR AW); and the Measure of Textual Lexical Diversity (MTLD Original) for all words.. Finally, lexical sophistication was also measured using TAALES 2.2 (Kyle & Crossley, 2018), focusing on indices derived from the Kucera-Francis (KF) and the British National Corpus (BNC). databases. These indices included: KF_Freq_AW (frequency of all words), KF_Freq_AW_log (logarithm frequency), KF_Ncats_AW (register range for all words), KF_nsamp_AW (range for all words), KF_Freq_CW (frequency of content words), KF_Freq_CW_Log (logarithm), BNC_Freq_AW (British National Corpus frequency all words), and the BNC_Freq_AW_log.

Lexical density, lexical diversity and lexical sophistication refer to statistical measures that gauge the lexical richness of texts. These indices may also be employed to assess student's progress, as it was done in this article to compare the developmental features of students across three cohorts.

Results and Discussion

The results and discussion are presented together in this section, addressing the constructs outlined in the research question: What developmental trends in lexical complexity – including lexical density, diversity and sophistication – can be observed by analysing learner writing from the CELTEC corpus across cohorts?

Lexical Density

To begin with, regarding the developmental trend of lexical density, the results in Table 4 reveal a gradual increase in the proportion of content (lexical) words across cohorts, indicating that students' writing becomes progressively denser with each grade (44,5% < 45.4% < 46.9%). Although this developmental trend is evident, these percentages are relatively low compared to students with similar proficiency levels (pre-intermediate to advanced). For instance, Zheng (2015) reported ratios close to 60% for Chinese advanced learners' writings, and Schnur and Rubio (2021) observed even higher results for Spanish learners with 60.66% of lexical diversity for lower proficiency (A1), and 92.72% for more advanced learners (B2). Zhang et al. (2021) reported a similar trend with a 57.6% ratio of lexical diversity.

In contrast, the findings of this article are different from the studies presented above. All the cohorts from CELTEC showed a lexical diversity ratio of around 45.5%, while the lower proficiency group from Zhang et al. (2021) reported ratios of 60.66% for A1 level and 92.72% for B2 learners. These findings suggest that CELTEC learners write simpler texts that are relatively easy to read, despite the steadily increasing information load from Year 3 to Year 5. The observed density is classified as average, typical of general prose, such as fiction and non-fiction, which combines varied vocabulary with function words. These results are in line with the Multidimensional analysis from the Biber Tagger indicating this corpus is closer to the general narrative text type.

Table 4: Lexical density in CELTEC across cohorts

Indices	Third year	Fourth year	Fifth year
Total number of tokens	355,4	599,00	739,07
Total number of content tokens	158,9	276,02	347,54
Lexical diversity	44.5%	45.4%	46.9%

Lexical Diversity

The developmental trajectory in lexical diversity, as presented in Table 5, reveals a substantial increase in indices across cohorts. All indices, except the Simple TTR for all words, exhibit marked growth between year groups. This progression demonstrates that learners produce longer texts (basic ntokens: 355.44 < 599.00 < 739.07) and employ a higher number of distinct word types (year 3 155, year 4 225, and year 5 263) as they advance in their IELTE programmes. Among these distinct word types, a significant proportion are lexical (content) types (105 < 161 < 194), outpacing the increase in function types (50 < 63.63 < 68.66). The trend is also reflected in lexical density types (0.66 < 0.69 < .071) and the Original Measure of Textual Lexical Diversity (63.6 < 67.43 < 68.6). These findings indicate that learners progressively extend their vocabulary, reducing repetition of the same lexical items over time.

This expanding vocabulary reflects students' ability to convey greater conceptual and informational content in their writing as they advance in their study programmes. By the higher levels (cohorts), learners exhibit improved lexical diversity, which aligns with Yu's (2010) assertion that lexical diversity serves as a critical measure for evaluating student writing. It also acts as a strong indicator of cognitive engagement and linguistic proficiency during the learning process. These results underscore the developmental gains in lexical richness as students progress through their education.

Nevertheless, the ratio of the TTR AW remains relatively low across all cohorts and does not show a clear developmental trend (0.48 > 0.43 > 0.41). This apparent contradiction may be explained by Johansson's (2008) observation that a lower TTR can result from the repeated use of common function words, particularly as text length increases. Johansson emphasises the importance of using more robust indices, such as the Measure of Textual Lexical diversity (MTLD), especially when analysing larger datasets, as was implemented in this study. This dual trend underscores the complexity of evaluating lexical diversity and highlights the necessity of employing multidimensional approaches to accurately capture linguistic development and growth.

Table 5: Lexical diversity in CELTEC across cohorts

Indices	Third year	Fourth year	Fifth year
Basic ntokens	355.44	599.00	739.07
Basic ntypes	155.82	225.19	263.22
Basic ncontent types	105.44	161.56	194.55
Basic nfunction types	50.39	63.63	68.66
Lexical density types	0.66	0.69	0.71
Simple TTR AW	0.48	0.43	0.41
MTLD_Original_ AW	63.6	67.43	68.6

Lexical Sophistication

Table 6 presents the results for the primary indices used to measure lexical sophistication in learner writing. These indices include the Kucera-Francis (KF) written frequency, based on the Brown Corpus. Despite the significant disparity in word counts among the three cohorts (28,335 for Year 3, 170,116 for Year 4, and 42,007 for Year 5), comparisons are feasible because the results from TAALES 2.2 are normalised to 1,000 words.

The analysis reveals no consistent developmental trend in lexical sophistication when comparing the learners' writing to the Brown Corpus. However, there is a gradual increase in sophistication when contrasted with the British National Corpus (BNC) written section (indices increase from 9.01 in Year 3 to 9.34 in Year 5). In contrast, the indices derived from the Brown Corpus such as the KF all words logarithm (log), display an irregular trajectory. For instance, the log frequencies of all words are higher for lower proficiency levels (year 3) than for higher proficiency levels (year 5) (3.044 > 3.025

> 3.015). Similarly, the FK frequency for content words reveals that the lower proficiency levels exhibit higher ratios than the upper levels (2.31 > 2.29 > 2.27).

These findings indicate that CELTEC learners do not demonstrate substantial developmental features of lexical sophistication across cohorts when measured against the Brown Corpus. It might indicate that the KF frequency, derived from a smaller and older corpus of written and spoken American English, is limited in its representativeness of contemporary usage.

In contrast, when using the BNC written section, which comprises 87,903,571 words of running text, a gradual developmental trajectory is observed. For example, the BNC_Written_Freq_AW index increases steadily (9.01 < 9.13 < 9.34), and the BNC_Written_Freq_AW_log shows a similar progression (-0.04 < -0.05 < -0.09). These results suggest that while CELTEC learners' writing exhibits limited development in lexical sophistication relative to older corpora, some gradual progress is evident when evaluated against more modern and extensive datasets.

Lexical Sophistication in CELTEC across cohorts

Indices	**Third year**	**Fourth year**	**Fifth year**
Word count	28,335	170,116	42,007
KF_Freq_AW	9845.043	10044.49	10167.44
KF_Freq_AW_Log	3.044	3.025	3.015
KF_Freq_CW	1099.554	1074.142	1069.793
KF_Freq_CW_Log	2.318	2.29	2.27
BNC_Written_Freq_AW BNC_Written_Freq_AW_log	9.01 -0.04	9.13 -0.05	9.34 -0.09

This comparison of datasets across year groups reveal that learners in higher cohorts are still prone to use high-frequency vocabulary in their writing. Their active use of low-frequency or more sophisticated words remains at a beginners' level, despite notable differences when comparing different corpora.

The results indicate that these learners' texts exhibit an average lexical density, which implies that the amount of information provided in the texts is relatively low. However, the cognitive load of the texts shows a steady increase with each academic year. Compared to similar learners in China and Spain, whose lexical density percentages approach 60% (Zheng, 2015; Schnur & Rubio, 2021), the CELTEC learners' average of 45.6%, which places them at a lower level.

Additionally, CELTEC learners show lexical diversity developmental features, with the number of tokens and types increasing each year, the same as the Measure for textual lexical diversity which is considered a much stronger index than the simple TTR for all words, whose ratios were rather low in this study.

Although these students evidence increasing lexical density and diversity in their writings, this does not mean that their texts deploy a more sophisticated use of vocabulary across cohorts, when compared to the Brown corpus. A slight difference is found if compared to a more modern corpus such as the BNC Written, where a moderate developmental trajectory can be observed.

Local learner corpora and learner corpus research

The building of local learner corpora is of paramount importance for educators, as it enables them to make informed decisions on materials development, lesson planning and feedback all grounded in the actual performance of the entire group. Without the use of corpora, teachers' conclusions are often based on partial data, limiting their ability to make comprehensive assessments. The creation of this type of learner corpus remains highly relevant, as McEnery et al. (2019, p.8) point out, "most

written learner corpora typically consist of essays produced by students. However, language users are required to deal with diverse EFL genres such as reports, proposals, and reviews" (see Table 2). These are just a few examples of writing forms that learners of English may need to master. Yet, these forms are often absent from the existing learner corpora. In this regard, the corpus presented in this study makes a significant contribution by including a more diverse set of genres, thus offering a more comprehensive representation of learner language in less represented EFL contexts.

CONCLUSIONS

This article presented the building and compilation of a local learner corpora using the Sketch Engine software, offering a clear, step-by-step methodology for replication. The corpus is representative of its context, consisting of written contributions from 139 learners across three cohorts enrolled in nine different universities in Chile. Additionally, it includes a variety of EFL genres beyond the typical essay, as encouraged by McEnery et al. (2019). In terms of lexical complexity, it is evident that they are making progress in vocabulary acquisition across cohors. However, their performance remains below that of similar learners in comparable contexts (Schnur & Rubio, 2021, Zhang et al, 2021; Zhang, 2022; Zheng, 2015). While measures of lexical density and diversity were relatively lower than those reported in previous studies, it is lexical sophistication that requires more attention. These students tend to overuse basic vocabulary at the expense of more advanced terms, a common feature of L2 learner writing.

In conclusion, these learners would benefit from continuous exposure to more challenging and complex input. Furthermore, incorporating different academic word lists, such as Coxhead's (2000) Academic Word List (AWL), could help enhance their lexical proficiency.

REFERENCES

Biber, D., Larsson, T., & Hancock, G. R. (2023). The linguistic organization of grammatical text complexity: Comparing the empirical adequacy of theory-based models. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2023-0016

Clavel-Arrotitia, B., & Pennock-Speck, B. (2021). Analysing lexical density, diversity, and sophistication in written and spoken telecollaborative exchanges. *Computer Assisted Language Learning Electronic Journal, 22(2), 230-250.* https://callej.org/index.php/journal/article/view/360/290

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66-79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford University Press.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S. (2004). Computer Learner Corpus Research: Current Status and Future Prospects. In U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics. A multidimensional Perspective*, 123-145. RODOPI.

Gray, B., Geluso, J., & Nguyen, P. (2019). The Longitudinal Development of Grammatical Complexity at the Phrasal and Clausal Levels in Spoken and Written Responses to the TOEFL iBT® Test. *ETS Research Report Series*, 1, 1-51. https://doi.org/10.1002/ets2.12280

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9-34). Cambridge University Press.

Halliday, M. A. K. (1985). An Introduction to Functional Grammar (1st ed.). London: Edward Arnold.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers*, 53, 61-79.

Johnson, 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.

Kyle, K., & Crossley A., S. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333-349. https://doi.org/10.1111/modl.12468

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

McCarthy, P. M., & Jarvis, S. (2010). MTLD, D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research Methods*, *42*(2), 381–392.

McEnery, T., Brezina, V., Gablasova, D. & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyse language use. *Annual Review of Applied Linguistics,* 39, 74-92. https://doi.org/10.1017/S0267190519000096

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Methods, Theory and Practice*. Cambridge University Press.

Paquot, M. (2018). Corpus Research for Language Learning and Teaching. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 359-374). Palgrave Macmillan.

Raufi, B., Mulyono, H., Puadi Ilyas, H., & Zulaiha, S. (2024). Exploring Indonesian EFL students' lexical diversity and its correlation with academic vocabulary use in an online academic writing environment, *Ampersand*, 13. https://doi.org/10.1016/j.amper.2024.100196.

Read, J., & Read, J. A. (2000). *Assessing vocabulary*. Cambridge University Press.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, *32*, 17-35. https://doi.org/10.1016/j.jslw.2016.02.002

Schnur, E., & Rubio, F. (2021). Lexical complexity, writing proficiency, and task effects in Spanish Dual Language Immersion. *Language Learning & Technology*, *25*(1), 53–72. https://hdl.handle.net/10125/73425

Wilkins, D. A. (1972). *Linguistics in Language Teaching*. Cambridge: MFT Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawaii Press.

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

Zhang, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics*, 3*2*, 371–396. https://doi.org/10.1111/ijal.12420

Zhang, X., Lu, X., & Li, W. (2022). Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres. *Applied Linguistics*, *43*, 168–195.

Zhang, H., Chen, M., & Li, X. (2021) Developmental Features of Lexical Richness in English Writings by Chinese Beginner Learners. *Frontiers in Psychology*, *12*. http://doi:10.3389/fpsyg.2021.665988

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English, *System, 56*, 40-53. https://doi.org/10.1016/j.system.2015.11.007.

.