BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Identifying Key Linguistic Variables of Second Language Speaking Proficiency Using Principal Component Analysis

Shinjae Park

Department of General Education, Kookmin University, Seoul 02707, Republic of Korea

ABSTRACT

This study conducted principal component analysis (PCA) to identify key linguistic factors distinguishing the speaking proficiency levels of Korean learners grouped by CEFR classifications. The PCA primarily focused on fluency, lexis, and complexity, with particular emphasis on syntactic complexity. The analysis revealed that PC1 explained the largest variance (31.6%) in proficiency levels, with syntactic complexity—measured as complex nominals per T-unit (CNT)—emerging as the most significant differentiating factor. An analysis of variance (ANOVA) confirmed statistically significant differences in CNT across proficiency levels, underscoring its critical role in evaluating grammatical ability. Fluency variables, such as articulation rate, and lexis variables, including lexical diversity, although statistically non-significant, provided valuable insights into learners' broader proficiency development. The results also indicated that higher proficiency levels were associated with greater use of syntactically complex structures, highlighting the importance of grammatical sophistication in second-language speech. Additionally, learners at lower proficiency levels relied more on simple sentence constructions, which suggests a gradual progression in syntactic complexity as proficiency increases. These findings emphasize the pivotal role of syntactic complexity in assessing speaking proficiency and suggest that incorporating such measures into language evaluation frameworks could provide a more comprehensive and nuanced reflection of learners' linguistic development. This, in turn, could facilitate the design of targeted instructional strategies and assessments that align more closely with learners' developmental trajectories, addressing specific gaps in their linguistic skills.

*CORRESPONDING AUTHOR:
Shinjae Park, Department of General Education, Kookmin University, Seoul 02707, Republic of Korea
Email: muhando@kookmin.ac.kr

ARTICLE INFO
Received: 29 October 2024 | Revised: 13 November 2024 | Accepted: 15 November 2024 | Published Online: 10 December 2024
DOI: https://doi.org/10.30564/fls.v6i6.7629

CITATION
Park, S., 2024. Identifying Key Linguistic Variables of Second Language Speaking Proficiency Using Principal Component Analysis. Forum for Linguistics Studies. 6(6): 623–633. DOI: https://doi.org/10.30564/fls.v6i6.7629

Keywords: L2 Speaking Proficiency; Syntactic Complexity; Lexical Complexity; Fluency; Principal Component Analysis (PCA)

1. Introduction

In the field of second language (L2) acquisition, understanding how different linguistic factors contribute to proficiency is a central area of research. Language proficiency is a complex construct influenced by various dimensions, such as fluency, complexity, and lexis. These dimensions encompass a wide range of language skills, including speech rate, sentence structure, and vocabulary richness. While these elements are substantially explored in written contexts, they may manifest differently in spoken language, which is often more spontaneous and less structured. The problem is that research on L2 learners' writing abilities is far more common than studies focusing on their speaking skills. This orientation is attributed to the fact that analyzing spoken language presents considerable challenges that are less prominent in the examination of written data.

One of the primary difficulties in spoken language analysis is the need for transcription. Unlike written text, spoken language includes hesitations, pauses, repetitions, and other phenomena that complicate the transcription process. Accurately converting speech into text requires careful attention to detail and a method that can capture the fluid nature of speech without losing important nuances. Even as these challenges exist, studying L2 speaking is crucial for comprehensively understanding a learner's overall language ability. Proficiency in speaking often reflects a learner's real-time linguistic processing skills and can reveal different aspects of language development that might not be as apparent in written language.

For such a purpose, a particularly well-suited approach is principal component analysis (PCA) because it allows for the reduction of a large set of variables into a smaller number of components that explain the most variance in data. This method not only helps identify the most significant variables contributing to proficiency differences but also ensures that these variables are grouped based on their shared contribution to underlying linguistic dimensions, such as fluency, complexity, and lexis.

Accordingly, this study conducted PCA to extract key variables that contribute to proficiency differences among Korean L2 learners of English. A range of variables related to fluency, complexity, and lexis were analyzed to pinpoint the most significant factors that explain variances between proficiency groups. This study also focused on how these variables, particularly those related to syntactic complexity, differ across three distinct proficiency groups. An analysis of variance (ANOVA) was carried out to determine whether these differences are statistically significant, thus providing deeper insights into the factors that best differentiate proficiency in spoken language. This research is important because it targets the L2 speaking domain, to which fewer studies have been devoted due to the aforementioned challenges. It was aimed at expanding the comprehension of how different linguistic dimensions contribute to overall language proficiency. The research questions that guided the investigation are as follows:

- Can PCA be used to identify the contributing variables within the key dimensions of language proficiency (fluency, complexity, lexis) that explain differences between proficiency groups?
- Do the key variables extracted through PCA show statistically significant differences among proficiency groups?

2. Literature Review

The CAF triad—complexity, accuracy, and fluency—has been widely adopted in research on L2 acquisition as a comprehensive framework for evaluating L2 learners' proficiency. These dimensions have been treated as focal issues in initial studies, but the past decade has witnessed an increasing recognition of the importance of lexis in assessing L2 proficiency. Scholars such as Skehan have emphasized that lexical richness and diversity are critical components of language ability and should be integrated into the CAF model [1]. This growing acknowledgment has led to a surge in studies that include lexical measures alongside complexity, accuracy, and fluency in the assessment of L2 learners.

Although the CAF triad enables the extensive evaluation of L2 proficiency, many studies have concentrated on a singular aspect of the triad in either the writing or speaking context. For example, Kuiken and Vedder focused on syntactic complexity in written tasks, observing that learners demonstrate greater accuracy and complexity in writing than in speaking due to the absence of real-time processing demands [2]. Using global measures such as clauses per T-unit, the authors found that task complexity significantly affects written complexity but does not produce similar effects in spoken activities. This contrast underscores how different modalities—writing versus speaking—affect the manner by which learners engage with each CAF component. That is, writing allows for greater complexity and accuracy than spoken language given the immediacy and fluency demands of the latter. These findings suggest that task type and modality should be considered carefully in L2 proficiency assessments covering the CAF dimensions. Conversely, De Jong et al. examined fluency in spoken L2 tasks and discovered that it is significantly influenced by task type and structure, with structured tasks yielding smoother, more controlled speech than open-ended tasks [3]. This highlights how fluency, as a distinct CAF component, changes uniquely depending on task demands, emphasizing the need for fluency-focused assessments in spoken language evaluation that are separate from complexity and accuracy considerations.

Numerous studies have also centered on complexity in L2 writing. Norris and Ortega, for instance, conducted a meta-analysis of how complexity, accuracy, and fluency are defined and measured across different contexts [4]. They found that complexity is often assessed through syntactic structures and that high complexity scores correlate with considerable proficiency. While Skehan and Foster [5] inquired into the cognitive demands of real-time processing in speaking, Kuiken and Vedder [2] explored similar dimensions in the context of written tasks, demonstrating that such tasks afford learners greater opportunities to emphasize accuracy and complexity compared with spoken activities. Building on this insight, Bulté and Housen investigated short-term changes in complexity during L2 writing development [6]. Their findings indicated that with focused instruction, learners can enhance the complexity of their output over time, suggesting that targeted teaching fosters complex language use in writing.

Contrastingly, studies focusing on spoken complexity are relatively scarce. Among the few initiatives is the work of Tavakoli et al., who explored the interplay between fluency and complexity in L2 speaking tasks [7]. Their findings suggested that learners who speak fluently also tend to exhibit complex language use, highlighting the interconnectedness of these dimensions in spontaneous speaking contexts. Similarly, De Jong et al. investigated how fluency, accuracy, and complexity manifest in different speaking tasks [8]. They emphasized the importance of task characteristics, implying that structured tasks generate different fluency profiles compared with open-ended speaking activities. To this observation, Biber et al. added that the real-time nature of speech production complicates the generation of highly complex structures, making it more challenging than writing [9].

Despite recent advancements, the analysis of L2 spoken data presents several inherent challenges. A primary issue is transcription, which requires accurately capturing not only spoken words but also the subtleties of hesitations and fillers, such as “ah,” “er,” “oh,” and “um,” that frequently occur in spontaneous speech [10]. Unlike written data, spoken language involves additional nuances that must be preserved to accurately reflect a speaker’s processing and communicative intentions. Beginner learners, for example, often leave pauses unfilled, resulting in disfluency, whereas advanced learners use fillers strategically, creating a more natural flow and making their speech sound more native-like [11].

Research has also indicated that hesitation markers, such as pauses and fillers, serve as indicators of cognitive processing difficulties, especially in high-demand tasks, and reflect a speaker’s linguistic planning. Silent pauses, although sometimes rhetorical, often reveal underlying processing challenges, affecting fluency and influencing perceived proficiency [12,13]. These findings stress that the analysis of L2 speech data should integrate detailed transcription practices to precisely capture the aforementioned markers, as they reveal valuable information about a speaker's language processing and proficiency.

Lexis, as an increasingly important aspect of proficiency research and as a component that encompasses both spoken and written language, has gained considerable attention in recent years. Crossley et al. analyzed lexical diversity and lexical sophistication in L2 speech, discovering that high-proficiency learners tend to utilize a broad and advanced range of vocabulary [14]. The authors punctuated the argument that lexical choices are key to demonstrating proficiency in conversational contexts. Kyle and Crossley also demonstrated that lexical sophistication, as measured by the frequency and complexity of words used, serves as a reliable indicator of proficiency in spoken contexts [15]. They further explored how learners of different proficiency levels employ sophisticated vocabulary and noted that high lexical sophistication in speech is often accompanied by context-appropriate word selection and nuanced language use. These characteristics reflect not only advanced vocabulary knowledge but also a learner's ability to navigate complex conversational needs, suggesting that lexical sophistication reliably distinguishes proficiency levels.

In a similar vein, Lu's extensive study based on oral narratives from the Spoken English Corpus of Chinese Learners highlighted important dimensions of lexical richness in L2 speech [10]. The author uncovered that lexical variation, rather than lexical sophistication or density, has the strongest correlation with raters' judgments of the quality of ESL learners' oral narratives. This suggests that lexical variation, which refers to the range of different words used, is a critical factor in evaluating speaking proficiency. These findings further imply that vocabulary instruction should prioritize increasing learners' lexical range over focusing on sophisticated vocabulary use. This conclusion aligns with the broader emphasis on diversity in lexical choice in spoken language research, reinforcing the idea that lexical variation is a more reliable predictor of oral proficiency than other measures of lexical richness.

Meanwhile, in the realm of written language, Read punctuated the critical role of vocabulary richness in evaluating L2 proficiency, noting that advanced learners consistently exhibit substantial diversity and sophistication in their written vocabulary [16]. Such studies have also indicated that lexical measures are vital for assessing proficiency, parallel to the dimensions of complexity and fluency, particularly when analyzing written tasks wherein learners have sufficient time to consider their lexical choices.

In summary, the existing body of research on the CAF triad has tremendously advanced our understanding of L2 proficiency, yet distinct patterns exist in the focus of studies on speaking versus writing. Fluency, complexity, and accuracy are often examined in isolation, with fluency predominantly explored in the context of spoken language, while complexity tends to be the focus in written language. While studies have addressed all three dimensions of the CAF triad in both speaking and writing settings, they highlighted the differing cognitive demands placed on learners in these two modalities.

Lexical analysis, meanwhile, is becoming an increasingly important aspect of proficiency research, with evidence indicating that both lexical richness and sophistication are integral to distinguishing proficiency levels in both speaking and writing. The present study built on this body of research by specifically focusing on L2 speaking, aiming to fill the gap in the literature by addressing the challenges of transcription and the treatment of fillers in spoken data. Employing PCA to analyze fluency, complexity, and lexis in L2 speech, this study sought to more deeply illuminate how these dimensions contribute to proficiency differences across learners.

3. Methods

3.1. Participants

This study analyzed data from the Incheon National University (INU) Multi-language Learner Corpus (MULC), which was compiled through speaking tasks administered to university students in South Korea. The dataset consisted of audio recordings from 187 students who voluntarily participated in response to public advertisements. These participants were randomly selected without consideration for major, gender, or age. For the purposes of this research, two-minute English monologues, conducted in a soundproof computer laboratory, were recorded in real-time. The participants were given a choice of four familiar and accessible topics to speak about. Given that most of them were freshmen who may have been inexperienced or anxious about speaking in English, the topics selected were those covering a broad range of issues, presented in the form of the following questions:

What do you usually do in your free time (hobbies, etc.)?

What is your favorite movie genre?

Do you believe true friendship can exist between two genders?

Is it better to have a dog or a cat?

The sample comprised 92 males and 95 females enrolled at a Korean university at the time of the study and had an average age of 20.9 years. The distribution of participants across proficiency levels, as assessed by a native-speaking professor according to the CEFR (**Table 1**), was as follows: 22 students at A1 proficiency and 42 at A2; 56 at B1 proficiency and 47 at B2; 18 at C1 proficiency and 2 at C2. For the analysis, the grouping was simplified into three distinct proficiency levels: A1 and A2 combined (64 recordings); B1 (56 recordings); and B2, C1, and C2 combined (67 recordings).

Table 1. Proficiency-based group composition.

Group 1 (64)		**Group 2 (56)**	**Group 3 (67)**		
A1	A2	B1	B2	C1	C2
22	42	56	47	18	2
Total 187					

3.2. Dimensions of Linguistic Proficiency

3.2.1. Syntactic Complexity Measures

The L2 Syntactic Complexity Analyzer (L2SCA) was used to evaluate the syntactic complexity indices implemented throughout the study. The L2SCA, developed by Lu, enables the automatic examination of syntactic complexity in L2 production [17]. It calculates various syntactic complexity metrics commonly used in language research. These metrics help researchers understand how complex the syntactic structures in L2 learners' productions are, which can indicate linguistic development and proficiency. Lu also categorized syntactic complexity indices into five major groups, along with the respective count of indices within each category: production unit length (three indices), overall sentence complexity (one index), the number of subordinate clauses (four indices), the number of coordinating clauses (three indices), and phrasal sophistication (three indices) [17]. This categorization advances the systematic assessment of the syntactic features of L2 writing, providing a structured framework for grasping how these indices contribute to overall syntactic complexity (**Table 2**).

Table 2. Syntactic complexity measures.

Variables	**Formulas**
Length of production unit	
Mean length of clause	MLC: words/clause
Mean length of sentence	MLS: words/sentence
Mean length of T-unit	MLT: words/T-unit
Sentence complexity	
Clauses per sentence	CS: clauses/sentence
Subordination	
Clauses per T-unit	CT: clauses/T-unit
Complex T-unit per T-unit	CTT: complex T-units/T-unit
Dependent clauses per clause	DCC: dependent clauses/clause
Dependent clauses per T-unit	DCT: dependent clauses/T-unit
Coordination	
Coordinate phrases per clause	CPC: coordinate phrases/clause
Coordinate phrases per T-unit	CPT: coordinate phrases/T-unit
T-units per sentence	TS: T-units/sentence
Particular structures	
Complex nominals per clause	CNC: complex nominals/clause
Complex nominals per T-unit	CNT: complex nominals/T-unit
Verb phrases per T-unit	VPT: verb phrases/T-unit

3.2.2. Lexical Complexity Measures

Lexical complexity comprises various dimensions and can be measured through specific metrics. A widely accepted framework identifies three main components: lexical sophistication/rarity, lexical diversity/variability, and lexical density [18,19]. In this study, the analysis of lexical complexity was guided by a comprehensive framework, and its various aspects were measured using the Text Inspector, an automated lexical analysis tool. The Text Inspector evaluates and provides detailed feedback on texts, analyzing key dimensions of lexical complexity, including lexical diversity, sophistication, and density [18].

Lexical sophistication was evaluated using the Academic Word List and Beyond-2000 scores based on the

British National Corpus (BNC) [20]. Researchers often use frequency-based metrics to compare lexical data in L2 production against language corpora [21], wherein the use of low-frequency vocabulary indicates progress in L2 production development [22]. Lexical diversity reflects the range of vocabulary in a text, measured through metrics such as MTLD and Vocd-D [23] (**Table 3**). Finally, lexical density, defined as the proportion of content words in a text [18], was determined using a manual process involving POS tagging and review to ensure accuracy.

Table 3. Lexical complexity measures.

	Variables	Definitions
Diversity	VOCD	A mathematical transformation of the standard type–token ratio (TTR), which reduces the intervening impacts of text length and indicates the degree of word repetition in a text
	MTLD	The average length of sequential word strings in a text that maintain a given TTR value
Density	Verbal E./S.	Verbal–word ratio
	Noun E./S.	Noun–word ratio
Sophistication	Beyond BNC (2K)	The Beyond-2000 values calculated by subtracting K1 and K2 ratios from 100%

3.2.3. Fluency Measures

Fluency in language production is a multifaceted construct that can be decomposed into several key factors: speed, composite measures, breakdown, and repair [3,24,25,26]. Speed refers to the rate at which a speaker produces speech, often quantified in words per minute. A high speech rate typically correlates with exceptional fluency, pointing to a speaker's capacity to produce language rapidly without excessive pauses. A composite measure encompasses various aspects of fluent speech, including speech rate, the frequency of filler pauses (e.g., "um," "uh"), and the overall smoothness of delivery, thereby providing a holistic view of fluency. Breakdown signifies interruptions in speech flow, such as hesitations or repetitions, which can disrupt the natural cadence of conversation and suggest low fluency levels. Finally, repair involves the strategies that speakers employ to correct errors or reformulate their statements, highlighting their proficiency in managing communication in real time. The method used to measure these factors in this work is described as follows [27]:

- Speed: Articulation rate, calculated as the total number of syllables divided by the total phonation time (excluding pauses) and multiplied by 60
- Composite measure: Speech rate (pruned), calculated as the total number of syllables divided by the total performance time (including pauses) and multiplied by 60
- Breakdown: Length of pauses per 60 seconds
- Repair: Frequency of all repairs per 60 seconds

3.3. Statistical Analyses

PCA and ANOVA were chosen as the methods for analyzing the data due to their complementary strengths in examining L2 speaking proficiency. PCA was performed to reduce the dimensionality of the data, allowing for the identification of underlying patterns and relationships among the various variables related to fluency, complexity, and lexis. By enabling the extraction of key components, the PCA facilitated a clearer understanding of how these dimensions interact and contribute to overall proficiency. Subsequently, ANOVA was applied to evaluate whether significant differences exist among the defined proficiency groups. This statistical approach allowed for the assessment of the impact of the identified components on proficiency levels, thus clarifying the relative importance of each dimension in L2 speaking performance.

4. Results

4.1. Principal Component Analysis

As previously stated, PCA was employed to reduce the data to a smaller number of principal axes when multiple independent variables exhibited correlations, thereby uncovering patterns among the variables. PCA was used for the following reasons: First, the variables related to fluency, complexity, and lexis have different ranges and units, which may potentially be correlated with one another. Analyzing these variables in their original form could make it difficult to clearly identify the individual effects of each variable. Therefore, PCA was adopted to summarize the variables, allowing for the identification of significant variance patterns and increasing the efficiency of the analysis.

Second, PCA enables the extraction of principal components that reflect correlations among variables while facilitating a better exploration of differences between proficiency groups. The derived principal components contribute to a clearer explanation of group differences and can serve as a foundational resource for additional statistical analyses of differences among proficiency levels.

4.1.1. Normalization

The variables used in this study consisted of diverse measures, such as fluency, complexity, and lexis, which have different units. Due to the varying ranges and units of these variables, a normalization process was necessary before conducting the PCA. If the scale differences between variables are unaddressed, the values of specific variables may appear disproportionately large, potentially skewing PCA results.

The normalization performed in this work involved the standardization of all the variables to have a mean of 0 and a standard deviation of 1. This process ensured that all the variables were on the same scale, allowing for a fair comparison of variables with different units in the PCA. The normalization was performed using the Standard Scaler and the equation below:

$$Z = (X–\mu)/\sigma \qquad (1)$$

where Z represents the standardized value, X denotes the original value, μ refers to the mean of the variable, and σ is the standard deviation of the variable. The PCA was then conducted, allowing for an equitable assessment of the contributions among variables and their effects on each principal component.

4.1.2. Principal Component Analysis

Table 4 summarizes the results of the PCA. The first principal component (PC1) accounted for 31.6% of the variance in proficiency levels, with the complexity variables (DCT, CT, VPT, etc.) identified as the primary contributing factors. The second principal component (PC2) explained 14.9% of the variance, with contributions from both complexity and lexis variables (MLC, VOCD, sophistication, etc.). The third principal component (PC3) accounted for 12.7% of the variance, with key contributing variables being those related to lexis and fluency (breakdown, repair).

Table 4. Summary of PCA results.

Principal Component	Variance Explained (%)	Key Contributing Factors (Top 5)
PC1	31.6%	DCT, CT, VPT, CS, CNT
PC2	14.9%	MLC, VOCD, sophistication, TS, speed
PC3	12.7%	VPT, MTLD, CS, breakdown, repair

The visual representation of the PCA results in Figure 1 allowed us to infer the explanatory power of the principal components on the basis of their distribution patterns. Although the figure does not explicitly provide eigenvalues or percentage variances explained, we could still draw some insights from the clustering and distribution of the data points. A wider range of distribution of points along the PC1 axis (the horizontal axis) compared with those appearing along the PC2 axis (the vertical axis) suggests that PC1 captured more variability in the data. This means that the differences among the groups were more pronounced in relation to PC1 than to PC2.

In this case, considerable dispersion along the PC1 axis in Groups 1 (purple in **Figure 1**) and 3 (yellow in **Figure 1**), while a tight distribution along the PC2 axis implies that PC1 accounted for a large proportion of the variance within the data. Therefore, even though the figure does not quantitatively indicate the variances explained, the visual spread reflects an initial qualitative assessment of the components' effectiveness in capturing the data's structure.

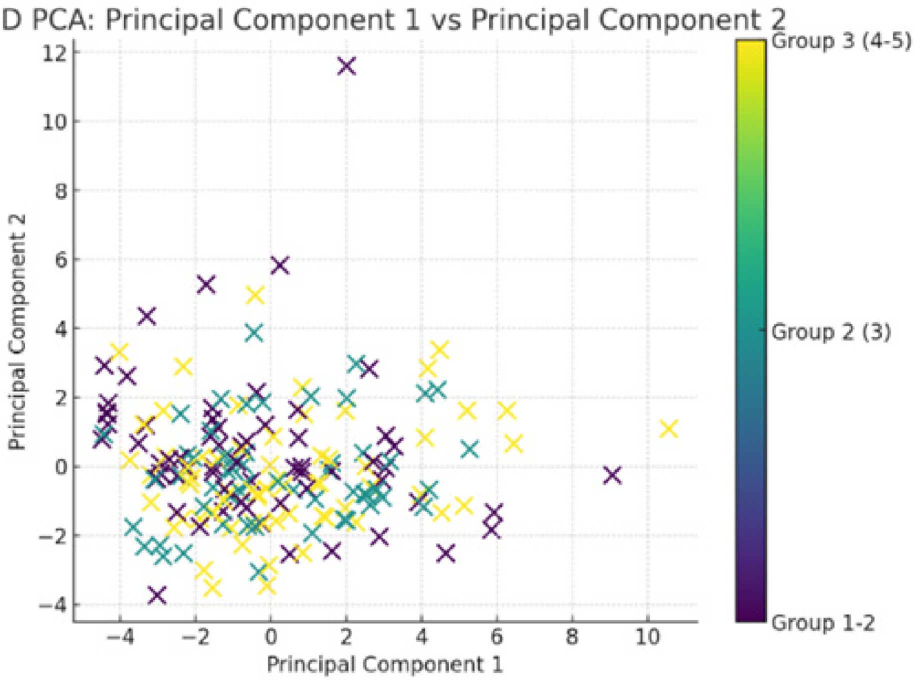

Figure 1. Principal component 1 vs. principal component 2 with group clustering.

Table 5 presents the variables that contributed most significantly to PC1. The complexity variables DCT and CT each exhibited a loading score of 0.34, indicating their cruciality in explaining the variance captured by PC1. VPT and CS also showed considerable influence on PC1, with their respective loading scores being 0.33 and 0.32. This pattern of results suggests that PC1 was predominantly characterized by complexity-related variables, which significantly contributed to the overall data variance. Collectively, these findings indicate that PC1 reflected the essentiality of complexity in understanding the underlying structure of the data.

4.2. Assessing Proficiency Differences by Grouping

Only PC1 served as the focal element because, first, a MANOVA analysis that covered PC1, PC2, and PC3 showed a statistically significant difference only in Roy's largest root (0.0528, p = 0.0252), whereas other criteria—specifically Wilks' lambda (0.9439, p = 0.1069), Pillai's trace (0.0564, p = 0.1082), and Hotelling-Lawley's trace (0.0591, p = 0.1063)—did not yield significant results. Second, PC1 explained 31.6% of the data variance, thereby justifying the decision to focus solely on this component in exploring the group differences.

Subsequently, ANOVA was conducted to scrutinize the differences among the proficiency groups (1–2, 3, 4–5) based on the PC1 factors (DCT, CT, VPT, CS, CNT) derived from the PCA. ANOVA was performed given that PC1 explained the largest proportion of the variance in the data, and it was deemed likely that the key variables contributing to this principal component would influence the differences among the groups. Additionally, it was necessary to statistically verify whether the characteristics of each proficiency group manifested differently in the variables contributing to the principal component (DCT, CT, VPT, CS, CNT).

Table 5. Summary of ANOVA results.

Variables	F-Statistic	p-Value
DCT	2.88	0.059
CT	2.35	0.099
VPT	2.76	0.066
CS	2.25	0.109
CNT	4.16	0.017

Table 5 presents the ANOVA results, which pointed to a statistically significant only in CNT (p = .017). This

variable was a crucial factor for explaining the differences among the proficiency groups, as it distinctly differentiated between the participants with high and low proficiency levels. The finding suggests that CNT reflected the characteristics of the learners who exhibited varying levels of proficiency. Conversely, the lack of significant differences in DCT, CT, VPT, and CS indicates that these variables did not contribute to the distinction in proficiency levels among the groups.

The findings also imply that the proficiency groups reflected CNT-related characteristics at a multivariate level. This may mean that certain complexity-related aspects captured by CNT are more pronounced in high-proficiency learners, whereas the other variables do not provide significant distinctions among groups.

5. Discussion

This study investigated the role of key variables within the CAF triad in distinguishing the proficiency levels of Korean learners of English. The analysis focused on syntactic complexity, particularly CNT, as it emerged as the most impactful variable in the differentiation. The emphasis on syntactic complexity aligns with the findings of Norris and Ortega [28], who highlighted that syntactic measures are essential in assessing language development, particularly in relation to productive skills such as writing. The current research built upon that insight by demonstrating that syntactic complexity can serve as a critical differentiator of spoken language proficiency—a context where real-time processing challenges render sophisticated syntax an indicator of advanced language ability [9].

This study likewise found that CNT was significantly effective in distinguishing among the proficiency groups, reinforcing the importance of syntactic complexity in L2 assessment. Previous research by Ortega [29] supports the idea that clause-based syntactic measures are associated with high proficiency, and the present study's findings add depth to this observation by confirming that such complexity is feasible and meaningful even under real-time constraints. This also aligns with Kuiken and Vedder's argument that syntactic complexity provides more consistent insights into proficiency than measures such as fluency or lexis, especially with respect to spontaneous spoken tasks [2].

Although CNT was identified as a statistically significant measure, the other variables, including DCT, CT, VPT, and CS, added value to understanding overall proficiency distinctions even though they did not reach statistical significance. This nuanced finding supports Ortega's observation that the utility of specific complexity measures can shift based on task demands and modality, implying that a multifaceted approach to L2 assessment is beneficial [29]. Lu similarly found that syntactic measures, such as clause length and subordination, correlate with proficiency in oral narration among Chinese learners, underscoring that multiple aspects of complexity can indicate proficiency beyond a single measure [10].

By centering the analysis on PC1, which captured the largest proportion of the variance in the data, this study foregrounded the value of a targeted principal component approach in L2 proficiency assessment. Ortega advocated for the selective use of complexity measures that capture core proficiency aspects across different tasks [29]. This focused approach aligns with the findings of Kuiken and Vedder, who contended that syntactic complexity is a stable and reliable indicator of proficiency, particularly in spontaneous speech contexts [2]. As proficiency levels in spoken language may be assessed with greater nuance by considering multiple syntactic measures, the present study supports the further exploration of targeted complexity indicators to improve accuracy in L2 assessment and deepen our understanding of language development.

6. Conclusions

The primary aims of this study were to examine differences in L2 proficiency among three distinct groups of learners by identifying key variables within the CAF dimensions of fluency, complexity, and lexis and to determine whether these variables show statistically significant distinctions across proficiency levels. To these ends, two research questions were pursued: Can PCA be used to identify the contributing variables within the key dimensions of language proficiency that explain differences between proficiency groups? Do the key variables extracted through PCA show statistically significant differences among proficiency groups? Spoken language samples were collected from learners who were classified into three CEFR-based proficiency groups, after which the data were processed through PCA to reduce the variable set while capturing the most variance.

The PCA results indicated that PC1, which explained the largest proportion of the variance in the data (31.6%), was the most relevant for distinguishing between proficiency levels. ANOVA tests on PC1 revealed that the proficiency groups exhibited statistically significant differences in CNT, representing syntactic complexity (p = 0.017). This result underscores the role of syntactic complexity, specifically CNT, as a reliable measure for differentiating proficiency, with high-proficiency learners demonstrating advanced syntactic structures. The other variables within the CAF dimensions, although statistically non-significant, contributed additional insights into the proficiency variations, supporting the multidimensional nature of L2 assessment.

This study confirmed that PCA effectively identifies key variables within the CAF framework, with syntactic complexity (CNT) as the primary distinguishing variable. The statistically significant differences in CNT across the proficiency groups emphasize the importance of including syntactic complexity measures in L2 proficiency assessments, particularly in spoken language contexts, to more accurately reflect learners' language development and competence.

Author Contributions

This study was conducted solely by the author, who took responsibility for all aspects of the research, including conceptualization, methodology, data analysis, writing, and project administration. The author strived to ensure the integrity and quality of the work throughout the process.

Funding

This work was supported by the National Research Foundation of Korea (NRE) (grant no. 2022S1A5A8054734).

Institutional Review Board Statement

The data used in this study was approved by the Institutional Review Board (IRB) at Incheon National University (IRB number: 7007971-201807-007A) under the authorization of the INU-MULC team.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

The data used in this study were obtained from the INU-MULC. The INU-MULC is publicly accessible, and more information about accessing the data can be found on its website (https://inu-mulc.inu.ac.kr). Restrictions may apply to certain portions of the data depending on licensing agreements.

Acknowledgments

The author would like to thank the anonymous reviewers for their valuable feedback and constructive comments, which helped improve the quality of the paper.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Skehan, P., 2009. Lexical Performance by Native and Non-Native Speakers on Language-Learning Tasks. In: Richards, B., Daller, M.H., Malvern, D.D., et al., (Eds.). Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application. Palgrave Macmillan, London. pp. 107–124.

[2] Kuiken, F., Vedder, I., 2012. Task Complexity and Linguistic Performance in L2 Writing and Speaking. Applied Linguistics. 33(2), 176–193.

[3] De Jong, N.H., Perfetti, C.A., 2011. Fluency Training in the ESL Classroom: An Experimental Study of Fluency Development and Proceduralization. Language Learning. 61(2), 533–568.

[4] Norris, J.M., Ortega, L., 2009. Towards an Organic

Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics. 30(4), 555–578.

[5] Skehan, P., Foster, P., 1997. Task Type and Task Processing Conditions as Influences on Foreign Language Performance. Language Teaching Research. 1(3), 185–211.

[6] Bulté, B., Housen, A., 2014. Conceptualizing and Measuring Short-Term Changes in L2 Writing Complexity. Journal of Second Language Writing. 26, 42–65.

[7] Tavakoli, M., Dastjerdi, H.V., Esteki, M., 2011. The Effect of Explicit Strategy Instruction on L2 Oral Production of Iranian Intermediate EFL Learners: Focusing on Accuracy, Fluency, and Complexity. Journal of Language Teaching & Research. 2(5), 989–997.

[8] De Jong, N.H., Steinel, M.P., Florijn, A., et al., 2012. Facets of Speaking Proficiency. Studies in Second Language Acquisition. 34(1), 5–34.

[9] Biber, D., Gray, B., Staples, S., 2011. Assessing Grammatical Complexity in Second Language Writing. Journal of Second Language Writing. 20(1), 2–23.

[10] Lu, X., 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal. 96(2), 190–208.

[11] Wang, Y.B., 2021. A Study on the Use of Hesitation Markers in Varied-Level EFL Learners' L2 Speaking Process. Open Journal of Modern Linguistics. 11(5), 824–834.

[12] Clark, H.H., Fox Tree, J.E., 2002. Using Uh and Um in Spontaneous Speaking. Cognition. 84(1), 73–111.

[13] Gilquin, G., 2008. Hesitation Markers Among EFL Learners: From Discourse Functions to Language Awareness. Corpora. 3(2), 147–172.

[14] Crossley, S.A., Salsbury, T., McNamara, D.S., 2014. Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy. Applied Linguistics. 36(1), 1–24.

[15] Kyle, K., Crossley, S.A., 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly. 49(4), 757–786.

[16] Read, J., 2000. Assessing Vocabulary. Cambridge University Press, Cambridge, UK.

[17] Lu, X., 2010. Automatic Analysis of Syntactic Complexity in Second Language Writing. International Journal of Corpus Linguistics. 15(4), 474–496.

[18] Bui, G., 2021. Influence of Learners' Prior Knowledge, L2 Proficiency and Pre-Task Planning on L2 Lexical Complexity. International Review of Applied Linguistics in Language Teaching. 59(4), 543–567.

[19] Tabari, M.A., Lu, X., Wang, Y., 2023. The Effects of Task Complexity on Lexical Complexity in L2 Writing: An Exploratory Study. System. 114, 103021.

[20] Park, S., 2023. Comparison of Lexical Complexity in L2 Speaking and Writing and Factors Predicting English Speaking Proficiency. 3L: Southeast Asian Journal of English Language Studies. 29(1), 125–138.

[21] Johnson, M.D., 2017. Cognitive Task Complexity and L2 Written Syntactic Complexity, Accuracy, Lexical Complexity, and Fluency: A Research Synthesis and Meta-Analysis. Journal of Second Language Writing. 37, 13–38.

[22] Laufer, B., Nation, P., 1995. Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics. 16(3), 307–322.

[23] McCarthy, P.M., Jarvis, S., 2007. Vocd: A Theoretical and Empirical Evaluation. Language Testing. 24(4), 459–488.

[24] Segalowitz, N., French, L., Guay, J.D., 2017. What Features Best Characterize Adult Second Language Utterance Fluency and What Do They Reveal About Fluency Gains in Short-Term Immersion? Canadian Journal of Applied Linguistics. 20(2), 90–116.

[25] Skehan, P., 2003. Task-Based Instruction. Language Teaching. 36(1), 1–14.

[26] Tavakoli, P., Nakatsuhara, F., Hunter, A.M., 2020. Aspects of Fluency Across Assessed Levels of Speaking Proficiency. The Modern Language Journal. 104(1), 169–191.

[27] Park, S., 2024. Analyzing Fluency Factors Across Speaking Proficiency Levels and Gender Among Korean English Speakers. Multimedia-Assisted Language Learning. 27(2), 60–70.

[28] Norris, J.M., Ortega, L., 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics. 30(4), 555–578.

[29] Ortega, L., 2012. Interlanguage Complexity: A Construct in Search of Theoretical Renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), Linguistic Complexity: Second Language Acquisition, Indigenization, Contact (pp. 127–155). De Gruyter Mouton, Berlin, Germany.