

Lexical and Syntactic Dimensions of Spoken English: A Comparative Analysis of ENS, ESL, and EFL Speaker Groups in ICNALE

YAMADA, Manako

School of Interdisciplinary Science and Innovation, Kyushu University

UCHIDA, Satoru

Faculty of Languages and Cultures, Kyushu University

<https://doi.org/10.15017/7410592>

出版情報：英語英文学論叢. 76, pp.1-20, 2026-03-18. Department of English, Faculty of Languages and Cultures, Kyushu University

バージョン：

権利関係：



Lexical and Syntactic Dimensions of Spoken English: A Comparative Analysis of ENS, ESL, and EFL Speaker Groups in ICNALE

Manako YAMADA,¹ Satoru UCHIDA²

1. Introduction

In Japanese English education, learners' low level of speaking proficiency has long been recognized as a critical issue. This problem is clearly demonstrated by large-scale national English proficiency surveys conducted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2018). The survey included about 60,000 third-year high school students nationwide, and approximately 10,000 were assessed for speaking. The results show that while 33.6% of learners reached CEFR (Common European Framework of Reference for Languages) A2 level or higher in listening, 33.5% in reading, and 19.7% in writing, speaking showed markedly lower attainment rates, with only 12.9% of learners achieving this level. Furthermore, the proportion of learners who received a zero score in speaking reached 18.8%, indicating that a substantial number of students were unable to demonstrate adequate oral proficiency. Notably, these results are not temporary fluctuations but rather consistent that closely resemble those of previous surveys. In other words, even after educational reforms that explicitly emphasized the importance of speaking skills were introduced, no substantial improvement in Japanese high school students' speaking proficiency has been observed. These findings underscore the need for research focusing on the speaking abilities of Japanese learners of English.

This study examines whether and how lexical and syntactic features differentiate English native speakers (ENS), English as a Second Language (ESL) speakers, and English as a Foreign Language (EFL) learners in spoken English, using spoken data from ICNALE. In addition, this study aims to clarify which of these features more strongly account for the observed intergroup differences.

1 School of Interdisciplinary Science and Innovation, Kyushu University

2 Faculty of Languages and Cultures, Kyushu University

Specifically, this study examines how ENS and ESL speakers—both of which can serve as reference models of target language systems for learners—differ from Japanese EFL learners in spoken English. By clarifying these differences through quantitative analyses of lexical and syntactic indices derived from ICNALE and the CEFR-based Vocabulary Level Analyzer (CVLA), this study seeks to provide a descriptive basis for understanding areas that may be relevant to the improvement of Japanese learners' English speaking proficiency.

The structure of this paper is as follows. Section 2 reviews previous studies on lexical and syntactic features in L2 speaking research, with particular attention to CIA-based approaches, research using ICNALE, and the use of quantitative lexical and syntactic indices. Section 3 describes the data, analytical indices, and methods employed in the present study. Section 4 presents the results of analyses based on lexical and syntactic indices, and Section 5 discusses these findings in relation to the research questions (RQs). Finally, Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

2.1 Learner Language Research and Contrastive Interlanguage Analysis (CIA)

In learner language research, learner production has not been regarded merely as an accumulation of errors. Instead, it has been widely conceptualized as interlanguage, a linguistic system that is distinct from both the learner's first language and the target language, yet internally systematic and analyzable. From this perspective, the characteristics of learner language should not be described in terms of what learners *cannot* do, but rather in terms of relative preferences in the use of linguistic forms and constructions (Granger, 1996).

This theoretical stance was methodologically operationalized through Contrastive Interlanguage Analysis (CIA), which involves systematic comparisons between learner groups and native speaker groups, or among different learner groups, under comparable conditions. CIA has been widely employed as a quantitative framework for identifying systematic features of learner language (Granger, 1996; Gilquin, 2008). While CIA has faced criticism concerning issues such as the comparative fallacy and the use of native-speaker norms, it has

since been reappraised as a methodology that explicitly addresses variability in learner language and enables more nuanced descriptions of interlanguage development (Granger, 2015).

2.2 CIA Studies on Speaking

In CIA-based research on speaking, a number of studies have examined differences between native and non-native speakers from various perspectives. For example, Götz (2015) analyzed fluency-related measures such as pauses and fluency-enhancing strategies including repetition and fillers. De Cock (2004) investigated recurrent sequences of words in the speech of native and non-native speakers of English. In addition, Diskin (2017) examined differences in the use of the discourse-pragmatic marker *like* between native and non-native speakers of English in Ireland, focusing on how second language speakers acquire both region-specific and pragmatic features. Despite these contributions, research that systematically investigates learner speech from lexical and syntactic perspectives using quantitative lexical and syntactic indices remains limited. Much of the existing work has focused on fluency, discourse markers, or formulaic sequences, rather than on the integrated analysis of lexical and syntactic characteristics.

In this regard, the International Corpus Network of Asian Learners of English (ICNALE) plays a crucial role in extending CIA-based approaches to research on L2 speaking (Ishikawa, 2013, 2018). ICNALE is a large-scale learner corpus that systematically collects spoken and written production from Asian EFL learners, ESL learners, and English native speakers (ENS) under strictly controlled and identical task conditions. By minimizing task- and condition-related variation, ICNALE enables robust intergroup comparisons that were difficult to achieve in earlier studies. In particular, the spoken module of ICNALE offers an empirical framework for investigating lexical and syntactic features of spoken language within a CIA perspective.

Previous studies using ICNALE have repeatedly reported biases in the lexical use in Japanese EFL learners' spoken production. For example, Suzuki (2023) examined the high-frequency verb *get*. The results showed that Japanese learners significantly overuse the *get* + *noun/NP* construction but underuse the

phrasal and prepositional uses of *get* that are common among native English speakers. This pattern suggests that the underlying issue is not a deficiency in lexical knowledge itself, but rather a lack of syntactic variety and flexibility in language use. In addition, Suzuki reported that the range of noun types occurring within the *get* + *noun/NP* construction was highly limited, indicating a lack of lexical expansion within a fixed construction. These findings suggest that lexical problems in Japanese learners' speech are not simply a matter of vocabulary size, but rather manifest as a fixation on specific lexical–syntactic pairings and a lack of flexibility in combining lexical items with constructions. Similar tendencies have been observed in studies focusing on multiword expressions. Sitler and Spivey (2025), using ICNALE data, reported that Japanese EFL learners' speech exhibits a restricted range of lexical bundle types and limited structural diversity, accompanied by a marked reliance on formulaic lexical–syntactic patterns.

On the other hand, analyses of syntactic features in spoken language have long been met with skepticism regarding the reliability of automated methods, due to issues inherent in speech transcription data. However, recent studies have demonstrated that automated syntactic analysis can function effectively even for L2 spoken data when appropriate preprocessing is applied. Kim and Lu (2024), for example, explicitly highlighted the importance of preprocessing in their analysis of syntactic complexity (SC) in L2 English speaking and empirically demonstrated that the L2 Syntactic Complexity Analyzer (L2SCA) can be used reliably. Their study showed that multiple syntactic indices—such as mean length of clause (MLC), complex AS-unit ratio (CAS/AS), dependent clause ratio (DC/C), and complex nominal ratio (CN/C)—consistently increased significantly with rising proficiency levels. These findings suggest that syntactic indices can serve as relatively stable indicators of L2 speaking proficiency. These syntactic characteristics manifest in learners' speech as differences in lexical output, lexical range, and structural elaboration. Taken together, previous studies indicate, first, that ICNALE provides a valuable data resource for conducting CIA-based speaking research by enabling comparisons among ENS, ESL, and EFL speakers under identical conditions; second, that the characteristics of Japanese EFL learners' spoken production should be understood not as a lack of

lexical knowledge but rather as biases in the use of lexical and syntactic resources; and third, that syntactic complexity indices, given appropriate preprocessing, are effective for capturing differences in speaking proficiency and learner groups.

2.3 The Need for an Integrated Examination of Lexical and Syntactic Features

When examining distinguishing differences in spoken production among ENS, ESL, and EFL speakers, the extent to which syntactic features and lexical features each contribute to speaker classification—and which of the two more strongly accounts for intergroup differences—has not yet been sufficiently examined. Although previous studies have demonstrated the usefulness of syntactic complexity indices and individual lexical features, relatively few studies have directly compared lexical and syntactic features within the same dataset and analytical framework to evaluate their respective discriminative and explanatory power. Recent research in computational language analysis has pointed out that language ability and lexical proficiency are multidimensional constructs that encompass not only lexical aspects but also syntactic and discourse-related dimensions (e.g., Crossley & Kim, 2022). Given the multidimensional nature of language ability, an analytical framework is required that can simultaneously handle multiple linguistic indices and uncover the latent structures underlying them. Accordingly, the present study adopts an integrative framework that jointly examines lexical and syntactic indices, under the assumption that the eight lexical and syntactic measures comprising CVLA3—originally developed to capture CEFR-J proficiency levels—also encode latent dimensions capable of reflecting intergroup differences in spoken production. Based on this framework, the present study addresses the following research questions.

RQ1: To what extent can ENS, ESL, and EFL speakers be clearly distinguished on the basis of lexical and syntactic indices in spoken production?

RQ2: What are the relative contributions of lexical features and syntactic

features to the classification of ENS, ESL, and EFL speakers?

3. Data and Methodology

3.1 Target Speaker Groups and Rationale for Data Selection

This study analyzes three speaker groups selected from the ICNALE corpus: English native speakers (ENS, $n = 300$), Philippine English speakers (PHL, $n = 200$), and Japanese EFL learners (JPN, $n = 300$). This categorization is grounded in Kachru's (1992) World Englishes framework, which conceptualizes English use in terms of a three-circle model consisting of the Inner Circle, the Outer Circle, and the Expanding Circle. As native speakers of English, ENS are classified as members of the Inner Circle and serve as a baseline for comparison. PHL speakers are situated in the Outer Circle, where English is widely used in institutional, educational, and public domains, and thus represent ESL speakers. In contrast, JPN speakers belong to the Expanding Circle, where English is primarily learned in classroom settings and opportunities for everyday English use are limited, characterizing them as EFL learners (Kachru, 1992).

By employing these three speaker groups, the present study explicitly operationalizes the ENS–ESL–EFL continuum based on differences in English use environments. This design enables a systematic examination of how lexical and syntactic characteristics in speaking relate to variations in English acquisition contexts. Although the analysis in this study is restricted to ENS, PHL, and JPN speakers, this limitation does not substantially compromise the representativeness of ICNALE. Uchida (2023), in an international comparative analysis using ICNALE, demonstrated that English use environment (ENS/ESL/EFL) functions as a major macro-level factor explaining variation in learner language. Cluster analysis revealed that ESL varieties (e.g., PHL, SIN, and HKG) exhibit patterns closer to those of native speakers, whereas many EFL varieties form distinct clusters separate from both ENS and ESL groups. These findings suggest that, while acknowledging diversity within the EFL category, the tripartite distinction of ENS, ESL, and EFL constitutes an effective analytical framework for capturing overall tendencies in learner language. Accordingly, treating JPN as a representative EFL group and PHL as a representative ESL group in the present study is consistent with previous ICNALE-based findings and appears

appropriate for examining intergroup differences based on English use environments. Therefore, although the present study does not encompass all countries and regions included in ICNALE, it appropriately selects three representative groups from the perspective of English use environments. As a result, the validity of the data foundation for investigating speaking differences among ENS, ESL, and EFL speakers is sufficiently maintained.

3.2 Task Control and Re-transcription of Speech Data

To control for task conditions, the present study analyzes both speaking tasks included in ICNALE; however, for each task, only the first speech sample produced by each speaker is used. This design aims to eliminate the effects of task familiarity and content planning that may arise from multiple speech productions by the same speaker, thereby allowing the analysis to more accurately reflect spontaneity and immediate language use in spoken performance. In the present study, the original transcripts provided in ICNALE were not used; instead, the speech data were newly transcribed from the audio recordings using an AI-based transcription system. The ICNALE spoken data contain many features characteristic of spoken language, such as hesitations, repetitions, false starts, and incomplete utterances, which are useful for qualitative, manual analyses. However, such features can introduce noise into automated syntactic parsing and the calculation of lexical indices. For this reason, the present study conducted re-transcription to ensure a level of formal uniformity suitable for quantitative analysis. The speech recognition model *gpt-4o-transcribe* was used for the transcription process. To preserve learner speech characteristics as much as possible, no corrections or improvements were made to grammatical errors, unnatural word choices, or non-native-like expressions; instead, the speech was transcribed verbatim, with the following prompt provided to the model.

PROMPT

This audio contains English spoken by learners. Please transcribe the speech as literally as possible. If there are grammatical errors, wrong words, or unnatural expressions, keep them exactly as they are and do not correct or improve them.

However, during the automatic transcription process, some cases of morphological-level normalization or minor corrections were observed. For example, utterances produced as *he experience* in the original audio were sometimes transcribed as *he experiences*, resulting in a grammatically regularized form. However, the focus of the present study is not on individual errors or error types themselves, but on the distributional patterns and structural tendencies of lexical and syntactic features. Therefore, such morphological overcorrections and minor normalizations were judged to have a limited impact on the overall analytical results, and the re-transcribed data were adopted for analysis. Moreover, the removal of noise caused by hesitations and similar phenomena enables more reliable automatic identification of syntactic units and more stable computation of lexical indices, which is considered appropriate given the aims of the present analysis. For these reasons, the present study conducted lexical and syntactic analyses based on ICNALE speech data that had been automatically re-transcribed using a unified set of criteria while also ensuring a relatively balanced data size across speaker groups.

3.3 CVLA Indices

Uchida and Negishi (2025), in a study focusing on reading data, reported that each of the indices constituting CVLA3 increases consistently as proficiency levels rise. Although the target modality differs from that of the present study, this finding indicates that lexical knowledge and lexical use are closely associated with language proficiency. As such, it provides important theoretical implications for analyses that focus on spoken production as well.

CVLA3 was developed as an extension of CVLA2 with the aim of more precisely capturing not only lexical complexity but also sentence- and construction-level complexity (Uchida & Negishi, 2025). AvrDiff is an index that calculates the average CEFR-J level difference of the words used, based on the CEFR-J wordlist, and reflects the difficulty level of lexical choices. BperA focuses on the amount of content word use and represents the ratio of B-level words to A-level words within a text. AvrFreqRank calculates the average frequency rank of words based on the Corpus of Contemporary American English (COCA) and reflects the degree to which the vocabulary used consists of low-frequency

words. CVV1 measures the dispersion of verb usage by calculating the coefficient of variation and reflects lexical diversity in verb selection, indicating whether speakers rely on a limited set of verbs or employ a wider range of lexical choices. ARI (Automated Readability Index) is a readability measure calculated from variables such as word length and sentence length. VperSent represents the number of verbs per sentence and reflects the complexity of sentence structure. POSTypes calculates the average number of distinct part-of-speech types per sentence and captures grammatical complexity indirectly through part-of-speech diversity. LenNP measures the average length of noun phrases functioning as subjects or objects and reflects the structural elaboration of major noun phrases. According to Uchida and Negishi (2025), AvrDiff, BperA, AvrFreqRank, and CVV1 are classified as indices of lexical complexity, whereas ARI, VperSent, POSTypes, and LenNP are considered indices of syntactic complexity. These indices were originally developed for the purpose of assessing the difficulty level of written texts, and their values generally increase in proportion to the text difficulty level. In the present study, it is assumed that these indices also exhibit meaningful level-related patterns when applied to spoken data. Based on this assumption, the study examines lexical and syntactic characteristics of spoken production by ENS, PHL, and JPN speakers from multiple perspectives. It should be noted, however, that the indices are not used to directly estimate CEFR levels in speaking, but rather as indicators of lexical and syntactic complexity.

3.4 Use of CVLA indices for Intergroup Comparison

In this study, lexical and syntactic indices based on CVLA3 were calculated for each speaker in the ENS, PHL, and JPN groups included in the ICNALE corpus. Only the raw values of the CVLA3 indices were used for analysis, and no level-converted values were employed (cf. Uchida & Negishi, 2025). Because learner data may contain anomalous values, multivariate outliers were identified using Mahalanobis distance. The upper 0.1% point of the chi-square distribution with degrees of freedom equal to the number of indices ($\alpha = 0.001$) was used as the threshold, and observations exceeding this value were classified as outliers ($n = 18$). Next, boxplots were generated using Matplotlib (version

3.10.7) to examine the distributional characteristics of each index and differences among speaker groups. In addition, the Shapiro–Wilk test was conducted for each index to assess normality. Because most of the indices (19 out of 24) did not satisfy the assumption of normality, non-parametric methods were used for testing statistical differences among speaker groups. Specifically, Kruskal–Wallis tests were conducted to examine group differences among ENS, PHL, and JPN, and for indices showing significant effects, post-hoc comparisons were performed using Dunn’s test with Bonferroni correction. Through these analyses, the presence of intergroup differences based on CVLA indices and their distributional characteristics was examined.

3.5 Factor-Analytic Procedures

Factor analysis was applied to the preprocessed lexical and syntactic indices to reduce redundancy among the CVLA3 measures and to examine how lexical and syntactic characteristics are organized into underlying latent dimensions. Because correlations among factors were assumed, Promax rotation, an oblique rotation method, was employed. The number of factors was set to three based on eigenvalues and the interpretability of the factor structure. Factor 3 showed relatively low loadings across all indices and was not included in subsequent analyses due to limited interpretability. Factor scores were computed for the extracted factors, and each speaker was plotted in a two-dimensional space defined by the first and second factors. This visualization made it possible to observe how the ENS, PHL, and JPN speaker groups were distributed, overlapped, or separated within the factor space. In addition, the composition of factor loadings was examined. Specifically, indices with high loadings on each factor were examined to determine whether they reflected lexical or syntactic characteristics, thereby clarifying the composition of each factor. This procedure made it possible to examine whether the latent dimensions contributing to intergroup separation were primarily lexical in nature or reflected more complex dimensions incorporating syntactic aspects.

4. Results

4.1 Intergroup Differences Based on CVLA Indices

This section provides an overview of intergroup tendencies among the three speaker groups—ENS, PHL, and JPN—based on lexical and syntactic indices derived from CVLA3. For each index, intergroup differences were examined using the Kruskal–Wallis test followed by post-hoc Dunn tests with Bonferroni correction, and the results are summarized in Table 1.

Table 1 Group differences in CVLA3 indices among JPN, ENS, and PHL speakers

	ENS	PHL	JPN	ENS-JPN	ENS-PHL	JPN-PHL
AvrDiff	1.561 (0.201)	1.606 (0.207)	1.411 (0.219)	$p < .05^*$	$p < .05^*$	$p < .05^*$
BperA	0.180 (0.103)	0.217 (0.134)	0.105 (0.116)	$p < .05^*$	0.057	$p < .05^*$
CVV1	2.348 (0.397)	2.202 (0.421)	1.581 (0.328)	$p < .05^*$	$p < .05^*$	$p < .05^*$
AvrFreqRank	378.906 (192.733)	423.419 (233.430)	349.655 (253.290)	$p < .05^*$	0.327	$p < .05^*$
ARI	11.352 (5.179)	11.245 (5.081)	6.940 (3.763)	$p < .05^*$	1	1
VperSent	4.885 (2.318)	4.343 (2.085)	2.980 (1.646)	$p < .05^*$	$p < .05^*$	$p < .05^*$
POStypes	9.638 (1.111)	9.525 (1.178)	8.174 (1.339)	$p < .05^*$	0.651	$p < .05^*$
LenNP	3.376 (1.503)	3.271 (1.357)	2.455 (1.325)	$p < .05^*$	1	$p < .05^*$

Notes:

- (1) An asterisk (*) indicates statistical significance at the $p < .05$ level; all p -values shown are Bonferroni-adjusted, with values exceeding 1.000 reported as $p = 1$.
- (2) Columns two through four present the mean values, with standard deviations shown in parentheses. All values are reported to three decimal places.

As shown in Table 1, the JPN group exhibited significantly different distributions from the ENS group across all indices, and from the PHL group across all indices except ARI. In contrast, significant differences between ENS and PHL were observed for only three indices: AvrDiff, CVV1, and VperSent. These patterns suggest that JPN speakers tend to be relatively distinct from the other

two groups across multiple aspects of lexical use, whereas differences between ENS and PHL are largely attributable to verb-related usage. In the following sections, two representative indices—CVV1, which reflects variability in lexical use, and VperSent, which captures syntactic expansion involving lexical elements—are examined in greater detail using graphical distributions.

4.1.1 Lexical Variability Index (CVV1)

Figure 1 illustrates the distribution of the lexical variability index CVV1 across the three speaker groups: ENS, PHL, and JPN. The median CVV1 value was highest for ENS, followed by PHL, while JPN showed a clearly lower value. The distributions of the three groups exhibited a stepwise difference, with JPN speakers tending to cluster toward the lower end of the scale overall. Focusing on the spread of the distributions, ENS and PHL showed relatively wide interquartile ranges, with observations extending into the higher-value range, indicating substantial individual differences in lexical variability. In contrast, the JPN group displayed a comparatively smaller degree of dispersion, visually confirming that lexical variability was more limited among these speakers. The results of the Kruskal–Wallis test indicated a statistically significant difference in CVV1 among the speaker groups ($p < .05$). Post-hoc comparisons using Dunn’s test revealed that the JPN group differed significantly from both the ENS and PHL groups in their distributions.

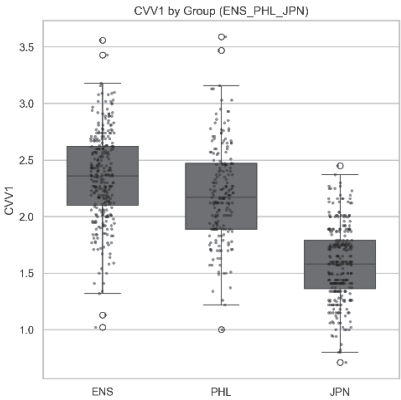


Figure 1 Distribution of CVV1 by speaker group

4.1.2 Indices Reflecting Syntactic Complexity (VperSent)

Figure 2 illustrates the distribution of VperSent, which represents the average number of verbs per sentence, across the speaker groups. Focusing on the

median values, ENS showed the highest VperSent, followed by PHL, while JPN exhibited lower values than the other two groups. As with CVV1, a stepwise difference was observed in the order of ENS, PHL, and JPN. Regarding the characteristics of the distributions, ENS and PHL showed relatively high values at the upper end of the scale, indicating that a certain number of utterances contained a relatively large number of verbs within a single sentence. In contrast, the JPN group showed an overall concentration toward lower values, suggesting a tendency toward fewer verbs per sentence and simpler sentence structures. The results of the Kruskal–Wallis test indicated that there was also a statistically significant difference among the speaker groups for VperSent ($p < .05$). Post-hoc comparisons revealed that the JPN group differed significantly from both the ENS and PHL groups in their distributions.

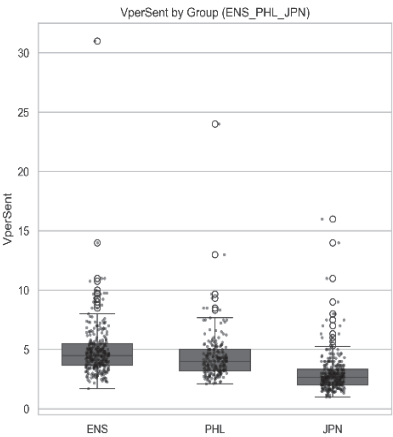


Figure 2 Distribution of VperSent by speaker group

4.2 Factor Interpretation

To examine the latent structure underlying the eight CVLA3 indices and to clarify how lexical and syntactic features contribute to intergroup differences, a factor analysis with Promax rotation was conducted. Based on the interpretability of the factor structure, a two-factor solution was adopted. Table 2 presents the standardized factor loadings for each index.

Table 2 Standardized factor loadings of CVLA3 indices for Factor 1 and Factor 2

Category	Index	Factor1	Factor2
vocab	AvrDiff	0.043	0.944
	BperA	0.039	0.886
	CVV1	0.434	0.212
	AvrFreqRank	-0.123	0.757
syntax	ARI	0.843	-0.025
	VperSent	1.147	-0.113
	POStypes	0.703	-0.124
	LenNP	-0.009	-0.003

4.2.1 Factor 1: Syntactic Complexity

Factor 1 showed high positive loadings for indices categorized as syntactic measures (Table 2). Specifically, VperSent (1.147), ARI (.843), and POSTypes (.703) loaded strongly on this factor, all of which reflect the structural elaboration of sentences through the number of verbs per sentence, readability, and part-of-speech diversity. In contrast, LenNP (−.009), which measures noun phrase length, showed virtually no loading on this factor. Among the vocabulary-related indices, CVV1 (.434) showed a moderate loading on Factor 1, suggesting that lexical variability in verb use may partially contribute to syntactic complexity. Based on these loading patterns, Factor 1 was interpreted as representing syntactic complexity in spoken production—specifically, the degree to which speakers construct elaborated sentences by deploying multiple verbs and diverse grammatical structures.

4.2.2 Factor 2: Lexical Complexity

Factor 2 showed high positive loadings for indices related to vocabulary difficulty and the proportion of content words (Table 2). AvrDiff (.944), BperA (.886), and AvrFreqRank (.757) loaded strongly on this factor, indicating that Factor 2 captures the level and sophistication of lexical choice. These three indices reflect different aspects of lexical complexity: the average CEFR-J level of words used, the ratio of B-level to A-level words, and the average frequency rank based on COCA. CVV1 also showed a moderate loading on Factor 2 (.212), indicating that verb diversity contributes primarily to syntactic complexity but also relates to lexical sophistication. In contrast, the syntactic indices (ARI, VperSent, POSTypes, LenNP) showed near-zero or negative loadings on Factor 2, confirming that this factor captures a dimension distinct from sentence-level complexity. Accordingly, Factor 2 was interpreted as representing lexical complexity—specifically, the difficulty level, frequency characteristics, and richness of the vocabulary deployed in spoken production.

4.2.3 Summary of Factor Structure

The factor analysis revealed a clear separation between syntactic and lexical dimensions. The syntactic indices (VperSent, ARI, POSTypes) loaded pre-

dominantly on Factor 1, while the lexical indices (AvrDiff, BperA, Avr-FreqRank) loaded predominantly on Factor 2. Notably, CVV1 showed moderate loadings on both factors (.434 on Factor 1; .212 on Factor 2), suggesting that verb diversity serves as a bridging variable that connects lexical and syntactic aspects of spoken production. LenNP, despite being categorized as a syntactic index, showed negligible loadings on both factors, indicating that noun phrase elaboration may operate independently of the dimensions captured by the other indices in this analysis.

4.3 Factor Analysis Plot

Figure 3 presents a two-dimensional scatter plot of individual speakers based on their Factor 1 and Factor 2 scores. The three speaker groups exhibited distinct distributional patterns in this factor space.

Along Factor 1 (syntactic complexity), a clear separation was observed between the speaker groups. The JPN speakers (circles) clustered predominantly in the negative region (left side of the plot), indicating relatively lower syntactic complexity characterized by fewer verbs per sentence and less diverse grammatical structures. In contrast, the ENS group (filled circles) was distributed broadly in the positive region (right side), with some speakers extending to values above 2.0, reflecting greater sentence-level elaboration and substantial individual variation. The PHL speakers (squares) occupied an intermediate position, with their distribution overlapping considerably with ENS but showing relatively less overlap with JPN. This pattern indicates that PHL speakers, as ESL users, exhibit syntactic complexity levels closer to those of ENS than to JPN.

Along Factor 2 (lexical complexity), all three groups showed considerable overlap across the entire range, from approximately -1.5 to 3.0 . Although group

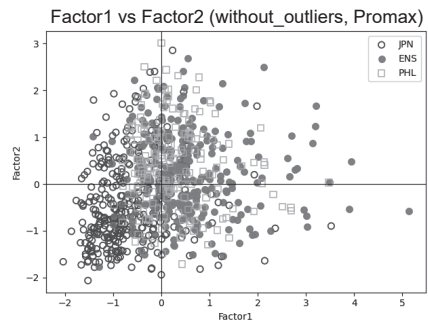


Figure 3 Scatterplot of speakers on Factor 1 and Factor 2

separation along this dimension was less evident than that observed for syntactic complexity, ENS and ESL speakers exhibited a broader distribution than JPN speakers, with values extending toward the higher end of the scale.

Notably, the central region of the plot (around the origin) showed substantial overlap among all three groups, particularly between ENS and PHL. However, JPN speakers were largely absent from the upper-right quadrant (high syntactic and high lexical complexity), whereas ENS and PHL speakers showed a spread into this upper-right region. This asymmetry suggests that the combination of high syntactic and lexical complexity is characteristic of ENS and ESL speakers and is rarely observed among JPN speakers.

5. Discussion

5.1 Separation of Speaker Groups Based on Lexical and Syntactic Indices (RQ1)

The first research question (RQ1) examined the extent to which ENS, ESL, and EFL speakers can be distinguished based on lexical and syntactic indices in spoken production. The results of Dunn's test with Bonferroni correction revealed that consistent and statistically significant differences were observed between the JPN and ENS/ESL groups across many indices analyzed (Table 1). These consistent differences suggest that the three speaker groups systematically differ in quantitative aspects of lexico-syntactic performance in speaking. In contrast, comparisons between the ENS and PHL groups showed a mixed pattern, with some indices revealing significant differences and others showing no significant differences. This pattern indicates that while PHL speakers share certain lexical and syntactic characteristics with the ENS group, they do not demonstrate native-like performance across all indices. In other words, PHL speakers, as ESL users, exhibited syntactic characteristics closer to ENS than to JPN, although they showed partial overlap with both groups. This pattern suggests that extensive English use in institutional settings may contribute to the development of syntactic complexity in speaking. In response to RQ1, the results indicate that ENS/ESL and EFL (JPN) speakers can be clearly distinguished based on lexical and syntactic indices.

5.2 Contributions of Lexical and Syntactic Features (RQ2)

The second research question (RQ2) aimed to clarify the relative contributions of lexical and syntactic features to the classification of ENS, ESL, and EFL speakers. The factor analysis yielded two distinct dimensions: Factor 1, reflecting syntactic complexity, and Factor 2, reflecting lexical complexity. Examination of the distribution in the factor space showed that ENS and JPN were relatively clearly separated along the direction of Factor 1, with PHL positioned between the two groups. These results suggest that differences in English use environments (ENS, ESL, and EFL) are reflected in distinct patterns of lexical and syntactic usage in spoken production.

Of particular importance is the fact that Factor 1 showed high factor loadings for indices incorporating syntactic elements such as *VperSent*, *POSTypes*, and *ARI*. This result implies that differences among speakers primarily reflect differences in sentence-level organization, particularly in how syntactic structures are used to integrate available lexical resources. In other words, differences in proficiency are likely to manifest as variation in speakers' ability to construct utterances through effective syntactic structuring. At the same time, because Factor 2 reflected indices related to lexical complexity, including the level of lexical choice and the amount of content words, lexical knowledge itself was also confirmed to be an important factor in explaining speaker differences. However, in the factor space, this factor alone did not clearly separate the speaker groups. Therefore, lexical knowledge is likely to exhibit stronger discriminative power when combined with lexico-syntactic deployment at the sentence level. These findings demonstrate that differences among ENS, ESL, and EFL speakers cannot be reduced to either lexical or syntactic features alone; rather, both the integrative deployment of lexical and syntactic resources at the sentence level and the level and quantity of lexical choice contribute in distinct ways. These findings carry implications for EFL pedagogy: instruction aimed at improving Japanese learners' speaking proficiency should address not only vocabulary size but also the flexible deployment of lexical items within varied syntactic structures.

6. Conclusion

This study investigated the extent to which lexical and syntactic indices account for differences among ENS, ESL, and EFL speakers in spoken production. Specifically, using multiple lexical and syntactic indices based on CVLA3, the study explored the latent structure of speech through factor analysis and examined the roles played by lexical and syntactic features in speaker classification. The results showed that ENS/ESL and JPN speakers were clearly separated along the syntactic complexity dimension, with JPN speakers clustering in the lower range. Along the lexical complexity dimension, group differences were less distinct than those observed for syntactic complexity, although some degree of differentiation among the three groups was still evident. Across both dimensions, PHL speakers tended to occupy an intermediate position, being distributed closer to ENS than to JPN. These findings suggest that differences in English use environments (ENS, ESL, and EFL) are reflected in patterns of lexical and syntactic deployment in speaking. The results of the factor analysis showed that lexical indices and syntactic indices loaded onto distinct factors, and that these two aspects constitute latent factors that differentiate the speaker groups.

Several limitations should be noted. First, the limited range of syntactic measures in CVLA3 may have constrained the analysis. Future research should incorporate more comprehensive syntactic complexity indices using tools such as L2SCA. Second, the applicability of CVLA indices to spoken data from the perspective of learner proficiency warrants further investigation. Third, within-group variation and comparative analyses with other EFL and ESL populations in ICNALE represent important directions for future research.

Use of AIs

In the preparation of this manuscript, ChatGPT 5.2 was used for translation and English-language proofreading. All content was reviewed thoroughly by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP23K21949 and JP24K00090.

References

- Crossley, S. A., & Kim, M. (2022). Linguistic features of writing quality and development: A longitudinal approach. *Journal of Writing Analytics*, 6, 59–93. <https://doi.org/10.37514/JWA-J.2022.6.1.04>
- De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *BELL: Belgian Journal of English Language and Literatures*, 12, 225–246. <https://dial.uclouvain.be/pr/boreal/object/boreal:75157>
- Diskin, C. (2017). The use of the discourse-pragmatic marker like by native and non-native speakers of English in Ireland. *Journal of Pragmatics*, 120, 144–157. <https://doi.org/10.1016/j.pragma.2017.08.004>
- Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: Detection, explanation, evaluation. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 1–33). Rodopi. https://doi.org/10.1163/9789401206204_002
- Götz, S. (2015). Fluency in ENL, ESL and EFL: A corpus-based pilot study. In Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015). Glasgow, UK.
- Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37–51). Lund 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>
- Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1, 91–118. <https://doi.org/10.24546/81006678>
- Ishikawa, S. (2018). The ICNALE spoken dialogue: Design of a spoken dialogue corpus for studies of L2 oral production. *Learner Corpus Studies in Asia and the World*, 3, 9–26. <https://doi.org/10.24546/81010115>
- Kachru, B. B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching*, 25(1), 1–14. <https://doi.org/10.1017/S0261444800006583>
- Kim, M., & Lu, X. (2024). L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic. *Modern Language Journal*, 108(1), 270–296. <https://doi.org/10.1111/modl.12907>
- Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2018). Gaikokugo kyōiku ni okeru “hanasu koto” “kaku koto” no hyōka no arikata ni tsuite [On the assessment of “speaking” and “writing” in foreign language education: Report of the expert panel for enhancing foreign language education]. https://www.mext.go.jp/a_menu/kokusai/gaikokugo/_icsFiles/afieldfile/2018/04/06/1403470_03_1.pdf
- Sitler, T., & Spivey, M. (2025). “I agree with this” : A multi-modal register analysis of lexical bundles in the ICNALE. *Asia Pacific Journal of Corpus Research*, 6(1), 25–45. <https://doi.org/10.22925/apjcr.2025.6.1.25>

- Suzuki, Y. (2023). Eigo gakushūsha hatsuwa kōpasu ni mirareru Nihonjin gakushūsha no kōhindo dōshi no shiyō: Dōshi get o rei ni [The use of high-frequency verbs by Japanese EFL learners in a spoken learner corpus: The case of get]. *Kyoiku Kaihatsu Kenkyu*. <https://doi.org/10.57529/00002075>
- Uchida, S. (2023). CEFR reberu ni yoru Eigo gakushūsha no sakubun no tokuchō bunseki: ICNALE ni motozuita chūkan gengo no kokusai hikaku. [Analyzing the Characteristics of English Learners' Writing by CEFR Level: An International Comparison of Interlanguage Based on ICNALE] In T. Otsu (Ed.), *Dēta o mochiita kotoba to komyunikēshon kenkyū no shuhō* [Methods for Research on Language and Communication Using Data]. (pp. 3–19). Hitsuji Shobo.
- Uchida, S., & Negishi, M. (2025). Estimating the CEFR-J level of English reading passages: Development and accuracy of CVLA3. *English Corpus Studies*, 32, 165–174. https://doi.org/10.69193/ecs.32.0_165