Asia Pacific Journal of Corpus Research
Vol. 2, No. 1, 35-45

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2021.2.1.35

Predicting CEFR Levels in L2 Oral Speech, Based on Lexical and Syntactic Complexity

Xiaolin Hu
(Tokyo University of Foreign Studies)

Hu, X. (2021). Predicting CEFR levels in L2 oral speech, based on lexical and syntactic complexity. *Asia Pacific Journal of Corpus Research, 2*(1), 35-45.

With the wide spread of the Common European Framework of Reference (CEFR) scales, many studies attempt to apply them in routine teaching and rater training, while more evidence regarding criterial features at different CEFR levels are still urgently needed. The current study aims to explore complexity features that distinguish and predict CEFR proficiency levels in oral performance. Using a quantitative/corpus-based approach, this research analyzed lexical and syntactic complexity features over 80 transcriptions (includes A1, A2, B1 CEFR levels, and native speakers), based on an interview test, Standard Speaking Test (SST). ANOVA and correlation analysis were conducted to exclude insignificant complexity indices before the discriminant analysis. In the result, distinctive differences in complexity between CEFR speaking levels were observed, and with a combination of six major complexity features as predictors, 78.8% of the oral transcriptions were classified into the appropriate CEFR proficiency levels. It further confirms the possibility of predicting CEFR level of L2 learners based on their objective linguistic features. This study can be helpful as an empirical reference in language pedagogy, especially for L2 learners' self-assessment and teachers' prediction of students' proficiency levels. Also, it offers implications for the validation of the rating criteria, and improvement of rating system.

Keywords: CEFR, Proficiency Predicting, Speaking Assessment, Lexical Complexity, Syntactic Complexity

1. Introduction

The Common European Framework of Reference for language learning, teaching and assessment (CEFR) (Council of Europe, 2001, 2018) has become such an influential standard for the development of language tests, curricula, educational standard and textbooks that nearly no (European) high-stake test was not related to it (Wisniewski, 2018). Also, CEFR scale system plays a crucial role to deliver the fair descriptions of learner language at different proficiency levels. For this purpose, specific criterial features are needed for CEFR's practical application in measuring L2 learners' performance.

To meet this need, the English Profile Programme (EPP) offers reliable information about which words (and importantly, which meanings of those words) and grammars are known and used by learners at each level of the Common European Framework. EPP focuses on vocabulary and grammars in the development of L2 language. However, it is not enough to describe and distinguish learners' proficiency adequately only through these two fields.

Besides vocabulary and grammar, it is more than necessary to investigate other linguistic features of L2 learners in the exploration of how L2 proficiency develops. Accordingly, there have been numerous attempts to identify various linguistic characteristics of L2 quality in terms of qualitative methods (e.g., Crossley, Salsbury & McNamara, 2012; Ferris, 1994; Jarvis et al., 2003; Lu, 2011). In these studies, various measures such as lexical density, lexical diversity, lexical variation, mean length of T-unit, and clause per T-unit have been widely used as major indices of L2 proficiency. Generally, complexity measures at lexical level and syntactic level are collectively referred to as lexical

complexity and syntactic complexity respectively. Among them, lexical complexity was confirmed to be a significant indicator of L2 learners' lexical and overall proficiency (Wolfe-Quintero, Inagaki & Kim, 1998). The other important aspect, syntactic complexity has been proved to be an important indicator especially of L2 writing (Crossley & McNamara, 2012; Lu, 2011, 2012). To build automated scoring models, some studies also attempted to explore syntactic complexity features as effective predictors to evaluate L2 learners' overall speech (Bhat & Yoon, 2015).

A clear understanding of how complexity measures contribute to overall speaking quality can hold significant implications for educators and researchers in terms of second language learning and teaching. Correspondingly, further analysis on predictors of CEFR proficiency become urgent demands. However, what specific lexical and syntactic indices can be combined together to distinguish different CEFR levels still remains unclear. To investigate complexity features as predictors of CEFR levels and supplement more Reference Level Descriptions (RLDs) for L2 oral performance, this study particularly focuses on L2 learners' lexical and syntactic complexity measures, with empirical data of eighty interview transcriptions from NICT JLE Corpus.

2. Background

2.1. CEFR and Criterial Features

As summarized in the Council of Europe's 2001 document, CEFR is intended to overcome the barriers of different educational systems in Europe, providing the common standard to situate L2 learners into six proficiency levels (Table 1).

Table 1. CEFR Levels

Proficient Users	C2	Mastery
	C1	Effective Operational Proficiency
Independent Users	B2	Vantage
	B1	Threshold
Basic Users	A2	Waystage
	A1	Breakthrough

In Chapter 3 of CEFR document, the Council of Europe proposed a large number of illustrative descriptors, which help to distinguish L2 learners of any languages between the six levels. In Appendix D (pp.244-57) it summarizes a set of "Can Do" statements developed by the Association of Language Testers in Europe (ALTE), which provides specific descriptions about what learners Can Do at different proficiency levels. Since this research discusses the complexity features in oral speech, only descriptors about speaking competence are listed in Table 2:

Table 2. The List of CAN Do Descriptors

ALTE Level	Council of Europe Levels	Listening/Speaking
ALTE Level 5	C2	CAN advise on or talk about complex or sensitive issues, understanding colloquial references and dealing confidently with hostile questions.
ALTE Level 4	C1	CAN contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good degree of fluency, coping with abstract expressions.
ALTE Level 3	B2	CAN follow or give a talk on a familiar topic or keep up a conversation on a fairly wide range of topics.

ALTE Level 2	B1	CAN express opinions on abstract/ cultural matters in a limited way or offer advice within a known area, and understand instructions or public announcements.
ALTE Level 1	A2	CAN express simple opinions or requirements in a familiar context.
ALTE Level Break-through Level	A1	CAN understand basic instructions or take part in a basic factual conversation on a predictable topic.

Table 2 shows that CEFR offers these descriptors in functional terms, which describe various functions L2 learners can perform when they gradually master a second language. In order to make it compatible with different languages, CEFR does not link any particular grammatical and lexical properties to itself. Consequently, CEFR levels are underspecified with respect to key properties, making it necessary for examiners to allocate candidates to a particular proficiency level in a particular L2 (Milanovic, 2009). The general descriptors about what learners can do does not illustrate with precision how the learner does it and with what grammatical and lexical properties of target languages (Hawkins & Buttery, 2010).

To ensure CEFR could be fully adapted to local contexts and purposes, the Council of Europe has encouraged the production of RLDs for national and regional languages. RLDs provide detailed, language-specific guidance for the use of the CEFR. Recently, various studies focus on the identification of lexical and grammatical "criterial features" for CEFR, most notably the EPP led by researchers from the University of Cambridge (Hawkins & Buttery, 2010; Hawkins & Filipovic, 2012).

2.2. Complexity Features

As complexity is a major aspect to measure L2 learners' performance, researchers have already widely used it in the proficiency prediction for written language. Many studies found it feasible to predict the proficiency level based on complexity features (Kim, 2014; Yoon, 2017). While due to the spontaneity of oral speech and the difficulty in its measuring, few studies made efforts to investigate the relationship between complexity features and proficiency levels in oral speech, and even less research focus on using complexity features for CEFR level predictions in speaking. Kang and Yan (2018) explored the linguistic features that could distinguish examinees' CEFR levels by a 1-minute-long monologic speech task, finding that complexity features are closely related to proficiency levels.

In current research, to describe complexity features in a more natural daily like context, instead of monologic speech, I evaluated transcriptions from an interview-based test, the Standard Speaking Test (SST). After summarizing their complexity features, a tentative predication was made for CEFR proficiency levels.

3. Methods

3.1. Data Set

As one of the world-largest spoken learner corpora, NICT JLE Corpus (Izumi, Uchimoto & Isahara, 2004) contains transcriptions of 1,280 L2 learners and 20 native speakers. The data of NICT JLE Corpus were collected from SST, an oral interview test for English in Japan, including five tasks within 15 minutes (Table 3). Each transcript has a score for proficiency levels (L1-L9). As defined by Tono (2013) (Table 4), proficiency levels can interconvert with each other based on the equivalent proficiency levels of SST, CEFR, and CEFR-J (the tailored CEFR for Japanese learning English). It should be noted that SST level 4 could be assigned to either CEFR level A1 or A2, while in this study, the evaluation of the classification method follows the later reallocating of CEFR-J Project, with SST level 4 included in the CEFR level A2. In this way, I reclassified the transcriptions into CEFR Levels, then collected eighty

samples, with twenty randomly selected from each of CEFR A1, A2, and B1, combined with twenty native speakers' transcriptions (Table 5).

Table 3. Tasks of SST

Stages of Exam	Question Type
1. Warm-up questions	Basic questions
2. Single picture	To describe the picture
3. Role-play with the interviewer	To read a card and role play with the interviewer
4. Picture sequences	To tell a story(based on given pictures)
5. Wind-Down questions	Casual questions

Table 4. Equivalent Levels CEFR, CEFR-J, and SST

CEFR	-	A1			A2		B1		B2		C1	
CEFR-J	Pre A1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.1	C1	C2
SST	1	2/3	3	4	4	5	6/7	8	9	9	9	9

Table 5. Number of Transcriptions in Each Group

CEFR Level	A1	A2	B1	Native Speaker
N	20	20	20	20

3.2. Linguistic Analysis and Automated Tools

Table 6 shows specific linguistic variables measured in this study. Lexical complexity was accessed by measures of lexical diversity, word length, lexical frequency and lexical density. Since the transcriptions are of quite different length, I used MTLD and VOCD to measure the lexical diversity for they are less influenced by text length. The number of tokens, MTLD, and VOCD were calculated by Text Inspector.

Table 6. Summary of Complexity Variables

	Submeasures	Descriptions
Lexical Complexity	Lexical diversity	Number of Tokens, Types, MTLD, VOCD
	Word length	Letters/word: number of letters per word Syllables/word: number of syllables per word
	Lexical frequency	Proportion of k1, k2, awl, and Off-List words
	Lexical density	Number of lexical words (or content words)/ the total number of words
Syntactic Complexity	MLT	Mean length of T-unit in words
	Clause/ T-unit (C/T)	Number of Clause per T-unit
	Coordinate phrase per T-unit (CP/T)	Number of Coordinate phrase per T-unit
	MLTurn	Mean length of Turn in words

Lexical frequency and lexical density were evaluated by the means of Web VocabProfile. This program calculated the proportion of words that were among the 1000 most frequent English word families (K1 words) and the 1000-2000 most frequent word families (K2 words). Meanwhile the percentage of words belonging to the Academic Word List (Coxhead, 2000) and the percentage of words that did not appear in any of the former lists were also computed (Off-List word). Additionally, Web VocabProfile offered lexical density results, by calculating the proportion of content words to the total number of words.

The syntactic complexity of speaking performance was accessed based on both general and specific measures. As a general index, a longer mean length of T-unit (the ratio of words to T-units) usually

indicates a higher complexity. Similarly, as it is an interview-based corpus and the interviewer and interviewee need to take turns during the speech, the mean length of turn (MLTurn) by interviewee draws a more overall portrait about how long one could speak for one turn in a conversation. For specific measures, this research applied the ratio of coordinate phrase to T-unit and clause to T-unit to investigate the syntactical complexity. The MLT, clause/T-unit, and coordinate phrase/T-unit were gauged by Web-based L2 Syntactic Complexity Analyzer (Lu, 2010). Though more syntactic complexity indices were accessible in this web-based Analyzer, after a small-scale pilot test, this research only adopted 3 of them for high reliability in L2 learner's oral speech.

3.3. Statistical Analysis

In data analyzing, the descriptive statistic describes basic patterns of the data. Then, after ANOVA examines the relationship between each feature and CEFR proficiency levels, several insignificant features were excluded. To investigate the correlations between different complexity features and avoid multicollinearity, this study also explored the correlations between each pair of the remained variables. Eventually, a discriminant analysis predicts the CEFR levels by complexity features. All the statistical analyses were performed on the Statistical Package for the Social Sciences (SPSS).

4. Result

4.1. Descriptive Statistic

Table 7 shows descriptive statistic for all complexity features tabulated by proficiency level. For lexical diversity, both of MTLD and VOCD show an increasing tendency along with the proficiency levels. As a method for lexical sophistication, mean length of words were calculated both in letters and syllables, while only letters/word visibly increases across levels. On the contrary, lexical density reveals a decrease tend.

Table 7. Complexity Features by Criteria and Proficiency levels

Criteria	Features	A1 (n=20)	A2 (n=20)	B1 (n=20)	NS (n=20)
		M(SD)	M(SD)	M(SD)	M(SD)
Lexical Complexity	N of Tokens	551.85(211.03)	1098.10(227.97)	1464.20(373.80)	4990.90(987.58)
	N of Types	182.55(40.56)	295.05(36.59)	381.60(63.19)	887.75(96.10)
	MTLD	21.93(6.70)	27.81(8.44)	33.79(8.89)	45.76(8.15)
	VOCD	51.83(12.24)	62.20(13.61)	69.24(13.66)	85.24(9.38)
	Letters/ word	3.43(0.23)	3.58(0.14)	3.69(0.13)	3.88(0.07)
	Syllables/word	1.25(0.06)	1.25(0.03)	1.25(0.03)	1.28(0.02)
	K1 words	63.05%(8.72%)	69.03%(6.70%)	78.81%(3.45%)	83.49%(2.76%)
	K2 words	5.00%(1.38%)	5.25%(1.42%)	4.85%(0.77%)	4.89%(0.75%)
Syntactic Complexity	AWL words	1.03%(1.14%)	0.93%(0.47%)	1.09%(0.66%)	1.03%(0.28%)
	Off-List words	30.92%(9.29%)	24.78%(7.37%)	15.26%(3.53%)	10.58%(2.37%)
	Lexical density	0.63(0.08)	0.57(0.05)	0.49(0.03)	0.46(0.02)
	MLTurn	8.28(3.75)	15.72(9.22)	19.50(7.36)	24.62(6.95)
	MLT	6.10(2.71)	7.91(1.78)	8.88(1.10)	10.71(1.17)
	Clause/ T-unit (C/T)	1.22(0.30)	1.44(0.23)	1.45(0.12)	1.61(0.16)
	Coordinate phrase/ T-unit (CP/T)	0.08(0.05)	0.11(0.04)	0.12(0.06)	0.19(0.05)

M=Mean; SD= Standard deviations

In cases of lexical resources (the proportion of K1, K2, AWL words, and Off-List words), the proportion of K1 negatively correlated with Off-List words. Interestingly, higher levels contain more

K1 words and less Off-List words. The main reason is that for basic users, lots of interjections like “erm, ee” were recognized as Off-List words (see example 1 from A1 level), and advanced and infrequent vocabularies had limited impact on results. Native speakers, on the other hand, could actually express themselves effectively with extremely basic words with much rare interjections. Another possible influence of interjections is that the value of letters/word is much lower for beginner L2 learners, since most interjections are much shorter than common words.

Example 1, A1- file00418,
A-interviewer, B-interviewee (Italicized words were recognized as Off-List Words)
(A: XXX02. Do you like movies?)
B: *Uum*. Yes.
(A: Um. So, please tell me about your favorite movie? The movie you like best.)
B: *Eeee*. I I like *eee* I like "Star Wars" best.
(A: What is the best scene in that movie?)
B: *Err*. *Ee*. I like Harison. *Ee. ee*.

Unlike lexical complexity, nearly all syntactic complexity features were notably and positively correlated with CEFR levels, and the discrepancy between native speakers and B1 level was obviously greater than the difference among A1, A2, and B1 groups.

4.2. ANOVA and Correlation Analysis of Predictor Variables

The ANOVA results of all 13 complexity features (Table 8) inspect whether each measure on its own showed statistically significant differences across 4 proficiency levels. Obviously, syllables/word, proportion of K2 words, and AWL words could be excluded for further analysis, since they have no overall significant effects across groups ($p > .05$). Additionally, about the predicting strength of each variable: the smaller the Wilks’ Lambda and the higher the F, suggests the greater the effect of the given measure. Therefore, the proportion of K1 word is most likely to be the strongest predictor for proficiency levels.

Then Table 9 investigates relationships between each pair of the ten variables, to avoid multicollinearity for further analysis. In this research r > .65 or r < -.65 was used as the cutoff value to leave out overlapped features. As expected, MTLD and VOCD were overly related with each other for both of them measures lexical diversity, and MTLD was remained for its stronger predicting strength. Similarly, proportion of Off-List words, lexical density, and clause/T-unit were excluded from the list of predictor variables. Interestingly, the lexical density was strongly negatively related with the proportion of K1 Words (r = -.865). Combined with information in Descriptive Statistics, as we mentioned the higher one’s proficiency level is, the more K1 words one would use in speech, and the lower the lexical density would be. The positive correlation between clause/T-unit and MLT (r = .891) was also predictable since a T-unit is likely to be longer when there are more clauses in it.

Table 8. ANOVA Output for All Predictor Variables

	Wilks’ Lambda	F	df1	df2	*p*
MTLD	.444	31.705	3	76	.000
VOCD	.495	25.828	3	76	.000
Letters/word	.449	31.051	3	76	.000
Syllables/word	.942	1.552	3	76	.208
K1	.342	48.793	3	76	.000
K2	.980	.506	3	76	.680
AWL	.994	.160	3	76	.923
Off-List	.374	42.471	3	76	.000
Lexical density	.349	47.300	3	76	.000

MLTurn	.574	18.787	3	76	.000
MLT	.529	22.584	3	76	.000
Clause/T-unit	.696	11.079	3	76	.000
CP/T-unit	.577	18.600	3	76	.000

Table 9. Bivariate Correlations of the Complexity Features

		MTLD	VOCD	Letters/ Word	K1	Off-List	Lexical Density	MLTurn	MLT	C/T	CP/T
Correlation	MTLD	1.000	.678	.301	.316	-.346	-.157	-.094	-.195	-.336	-.015
	VOCD	.678	1.000	.489	.534	-.577	-.334	-.188	-.108	-.205	.080
	Letters/word	.301	.489	1.000	.349	-.431	-.043	-.058	-.148	-.198	.182
	K1	.316	.534	.349	1.000	-.972	-.865	-.049	-.057	-.151	.075
	Off-List	-.346	-.577	-.431	-.972	1.000	.817	.085	.063	.167	-.091
	Lexical Density	-.157	-.334	-.043	-.865	.817	1.000	.017	-.056	.026	-.044
	MLTurn	-.094	-.188	-.058	-.049	.085	.017	1.000	.265	.363	-.023
	MLT	-.195	-.108	-.148	-.057	.063	-.056	.265	1.000	.891	.618
	C/T-unit	-.336	-.205	-.198	-.151	.167	.026	.363	.891	1.000	.425
	CP/T-unit	-.015	.080	.182	.075	-.091	-.044	-.023	.618	.425	1.000

4.3. Discriminant Analysis

Based on the lexical and syntactic features, a discriminant function was conducted to predict the CEFR levels and native speakers. After removing the insignificant and over correlated measures, the discriminant analysis includes 6 predictor variables: MTLD, letters/word, proportion of K1 words, lexical density, coordinate phrase/T-unit, MLT, and MLTurn. Multiple assumptions for data quality were met, and the relatively large sample size (N=80) as well as sufficient within-group sample size (n=20) suggested that the analysis would be robust. A test for equality of group means indicated statistically significant ($p < .05$) differences between the four groups on each of the six variables. The bivariate correlations between all pairs of variables ranged from -.195< r < .618, indicating that though there are some overlapping relationships here, none is strong enough to suggest multicollinearity (r< .65). Therefore, all variables on their own predicted CEFR levels and were retained for discriminant analysis.

The analysis identifies three discriminant functions, the first accounting for the large majority (94.8%) of observable variance across the four proficiency levels. An overall statistically significant effect was found for the combined three functions, Wilks' lambda = .139, χ^2 (18, N=80) = 146.031, $p < .001$, indicating that 86.1% of the actual variance among four proficiency levels could be explained. Both the second and third function did not provide additional statistically significant predictions. As shown in Table 10, Function 1 was best represented by the proportion of K1 words, correlated at .632, followed by letters/word and MTLD. In general, compared to syntactical complexity, lexical complexity plays a more important role in Function 1.

Table 10. Structure Matrix

	Function		
	1	2	3
Letters/word	.511*	.131	.046
MTLD	.510*	.367	-.263
MLT	.435*	.070	.380
CP/T-unit	.373	.611*	-.124
K1	.632	-.402	-.647*
MLTurn	.394	-.112	.515*

*. Largest absolute correlation between each variable and any discriminant function

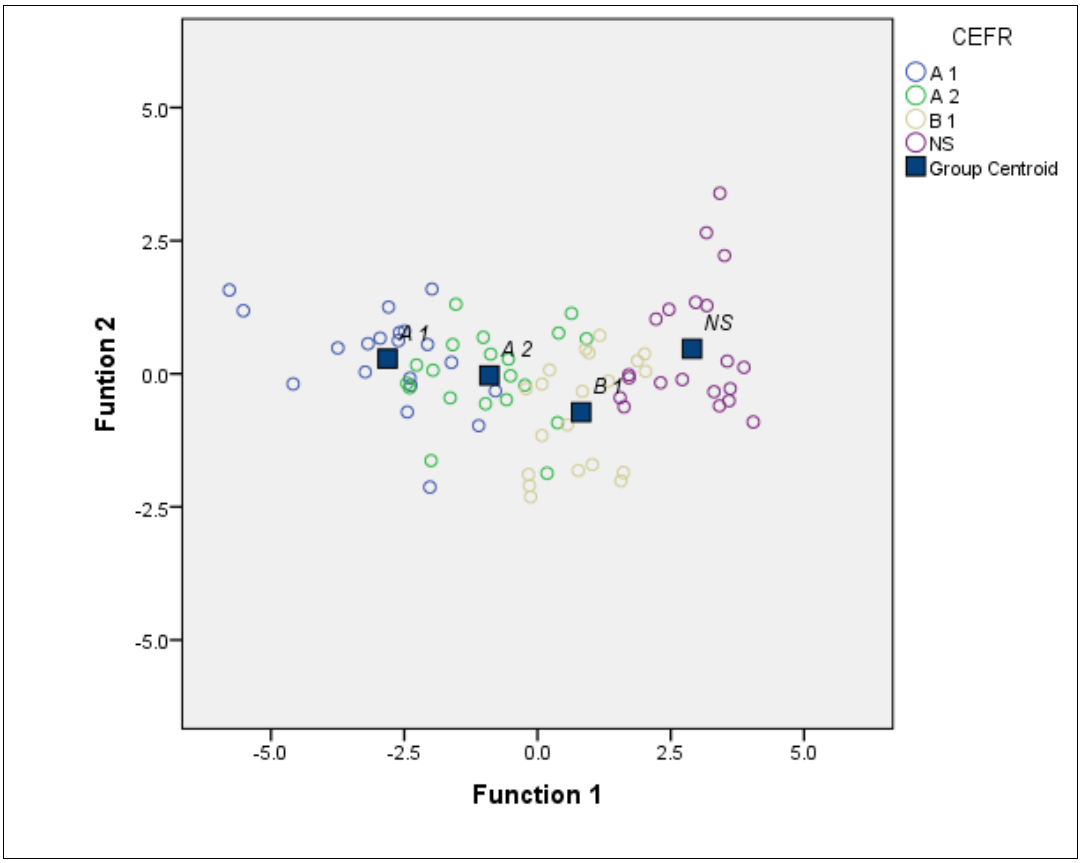

Figure 1. Canonical Discriminant Functions

Figure 1 shows the individual cases and group centroids (average values for each proficiency) displayed in two dimensions for the best two functions. Function 1 clearly distinguishes between all four proficiency levels, and slightly more so between native speakers and other CEFR levels. Nonetheless, several plots of native speakers were notably distributed closer to B1 level, leading to a decrease in its classification accuracy. Function 2 additionally distinguishes between B1 and native speakers, but less so between the other groups.

Finally, Table 11 represents the classification results for the discriminant analysis. Over all, the combination of three functions were able to classify 78.8% of the transcriptions correctly into their proficiency levels. Not surprisingly A2 group was the thorniest one to predict, with only 60% of them were classified correctly. Intriguingly, A1 and B1 levels were actually easier to predict than native speakers. Combined with the information from Figure 1, though the mean discrepancy between native speakers and B2 seems to be larger, a minority of native speakers tend to speak in a similar way with B1 level in lexical and syntactic complexity, leading them to be more difficult to be discriminated only by complexity features.

Table 11. Classification Results

			Predicted Group Membership			
		CEFR	A 1	A 2	B 1	NS
Original	Count	A 1	18	2	0	0
		A 2	3	12	5	0
		B 1	0	0	17	3
		NS	0	0	4	16
	%	A 1	90.0	10.0	.0	.0
		A 2	15.0	60.0	25.0	.0
		B 1	.0	.0	85.0	15.0
		NS	.0	.0	20.0	80.0

a. 78.8% of original grouped cases correctly classified.

5. Discussion

As a preliminarily attempt, the current study illustrates complexity features in speaking performance that could predict CEFR levels and native speakers. The overall result suggests that the distinctive difference in several complexity features is capable to classify at least part of the oral performance correctly.

According to the result in this research, compared to syntactic complexity, lexical complexity features were more closely related to proficiency levels. As the key predictor variable, K1 words dramatically increased along with proficiency levels. It prompts us that compared with the expansion of advanced vocabulary, it might be more helpful to express oneself with basic words in a more natural way. Meanwhile the increase in lexical diversity and letters/words could positively affect the proficiency predicting as well. Similar tendency was also validated for L2 learners' monologic speech in Kang and Yan's research (2018).

There are some limitations to this study that need to be acknowledged. First, limited by the deficiency of adequate automated tools for L2 learners' oral speech measuring, some sophisticated features of syntactic complexity were not contained in this research. Consequently, the result might be inadequate to capture inclusive characters for syntactic complexity. However, though it is not a comprehensive covering, features like MLT and Coordinate phrase/T-unit did sketch a general silhouette for both L2 learners and native speakers in syntactic complexity. Moreover, in the current study, three syntactic features were found to be statistically significant to discriminate L2 learners between adjacent CEFR levels.

Another limitation is that it seems not fair enough to predict the proficiency levels only based on complexity features. While as complexity is one of the major parts for speaking competence, it might be a wise choice to start with complexity features. Though the classification result might be slightly different if native speakers' group is not included, the effective predicting strength of complexity features for CEFR was still clearly shown in this study. In future research it would be interesting to explore the CEFR level prediction with more linguistic features included, such as fluency, accuracy, and pronunciations.

Nevertheless, despite all above limitations, the findings in this study have yielded valuable new insights on CEFR prediction by objective linguistic features, and confirmed the significance of lexical and syntactic complexity in L2 spoken language.

6. Conclusion

As an exploratory study, the research aims to identify complexity features that could predict CEFR levels based on oral performance in the interview-based test. The result of the study made two things explicit: a) six complexity features useful for the prediction of CEFR levels; and b) the actual variations of complexity features among different CEFR levels and native speakers.

These findings lead to meaningful implications for L2 learners, educators, and raters. First, lexical and syntactic complexity could be considered as useful factors in L2 learners' self-assessments, examining in what aspect one should particularly make improvements to get into higher levels (e.g., diversity of vocabulary, syntactical structure, and etc.). Second, the method proposed in this study may make it easier for teachers to perform suitable teaching methods for different groups (levels), since the prediction reveals that it is possible to use automated tools to classify students into different groups based on their linguistic features in oral performance. Especially for speaking class, where level-specific instruction is imperative for students (Bailey, 2005), the differences between CEFR levels could be designed as classroom tasks to help students make progress in lexical and syntactical complexity. Third, for the rating system of spoken language, this study can be used as a useful reference to

supplement more criterial features of L2 learners. Consequently, it may be helpful for the further application of CEFR, especially in those CEFR aligned tests.

Concerning the limitation of this study, it only considers part of syntactic complexity indices due to the lack of adequate tools for spoken language. Future studies will benefit from adding more syntactic indices with more developed tools. Also, though this study only focused on complexity, it verifies the possibility to capture differences between CEFR levels by linguistic features, especially with complexity indices. It will be a worthwhile future research direction to predict CEFR levels from a thorough analysis, including fluency, accuracy and etc.

In summary, more attention should be paid on exploring the specific criterial features of CEFR levels, which may improve the rating system of L2 learners' spoken language and shed lights on the applications of objective linguistic features in CEFR.

References

Bailey, K. (2005). *Practical English Language Teaching: Speaking*. New York: McGraw-Hill.

Bhat, S., & Yoon, S. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67*, 42-57.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with New Descriptors.* Strasbourg: Council of Europe.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

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.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243-263.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Hawkins, J., & Buttery, P. (2010). Criterial features in learner corpora: theory and illustrations. *English Profile Journal, 1*(1).1-23.

Hawkins, J., & Filipovic, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework.* Cambridge: Cambridge University Press.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management ,12*(2), 119-125.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment, 1,* 24-39.

Kim, J. (2014). Predicting L2 Writing Proficiency using linguistic complexity measures: A corpus-based study. *English Teaching, 69*(4), 27-51.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4),474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The*

Modern Language Journal, 96(2), 190-208.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes, 37,* 2-5.

Tono, Y. (Ed.). (2013). *The CEFR-J Handbook: A Resource Book for Using CAN-DO Descriptors for English Language Teaching.* Tokyo: Taishukan Publishing.

Wisniewski, K. (2018). The empirical validity of the Common European Framework of Reference Scales. An exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics, 39*(6), 933-959.

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

Yoon, H. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality, *System, 66*, 130-141.

THE AUTHOR

Xiaolin Hu is a PhD student at Tokyo University of Foreign Studies. Her principal research lies in the field of corpus linguistics and identification of CEFR criterial features.

THE AUTHOR'S ADDRESS

First and Corresponding Author
Xiaolin Hu
PhD Student
Graduate School of Global Studies
Tokyo University of Foreign Studies
3-11-1, Asahi-cho, Fuchu-city, Tokyo 183-8534, JAPAN
E-mail: hu.xiaolin.t0@tufs.ac.jp

Received: 29 October 2020
Received in Revised Form: 30 May 2021
Accepted: 28 July 2021