「論文」

Exploring L2 Spoken Developmental Measures: Which Linguistic Features Can Predict the Number of Words?

Yuichiro KOBAYASHI, Mariko ABE, and Yusuke KONDO

Abstract

One of the challenges for research in second language (L2) acquisition is finding reliable indices to objectively measure language development. To this end, researchers usually compare language learners of different proficiency levels through language proficiency tests. However, these proficiency levels can vary because each proficiency scale has different objectives and evaluation criteria. If the levels to be compared change, the developmental indices identified in the comparison change accordingly. Considering these issues, we seek to explore the effectiveness of criteria other than test scores and proficiency levels. Statistically, word tokens can be an alternative measure of spoken proficiency levels, as there is a high correlation between speaking proficiency and the number of words used in L2 speech. In addition, word tokens can be measured objectively and more consistently than proficiency levels. The number of words need not be converted from test scores, as it can be directly calculated from learners' spoken performance. Given these advantages, the present study investigates the mechanism of the increase in word tokens in L2 speaking. To do this, we counted the frequencies of Biber's (1988) 67 linguistic features in 832 L2 speech samples. Using these frequencies as predictor variables for random forest regression analysis, the study identified the features that contribute to an increase in the number of words. The results suggest that (a) causative adverbial subordinators, (b) independent clause coordination, (c) emphatics, (d) nouns, (e) prepositional phrases, and (f) present tense can best predict language development. These six key features can be robust indices of spoken language progress because they are frequently used in almost all speaking situations. The findings of the current study also offer valuable new insights into the methodology of L2 developmental studies.

『英語コーパス研究』第29号（2022），pp. 1–18

1. Introduction

Indices that effectively and objectively measure language development are of notable challenge in the field of second language acquisition (SLA) research. To tackle this challenge, researchers have compared various linguistic characteristics, such as lexical diversity and syntactic complexity, extracted from performances by learners with different proficiency levels (Crossley & McNamara, 2012; Díez-Bedmar & Pérez-Paredes, 2020; Kyle et al., 2021; Kyle & Crossley, 2018; Lu, 2011; Tracy-Ventura et al., 2021; Verspoor et al., 2021; Vyatkina, 2013). However, the estimated proficiency levels based on the scores of different language tests can vary because different language tests have different objectives and evaluation criteria. For example, in a particular test, the number of grammatical errors is a good predictor of the estimated proficiency levels; however, in another test, it can be a poor predictor of the proficiency levels. Considering this issue, the effectiveness of criteria other than test scores and proficiency levels must be explored.

Word tokens have shown promise as an alternative measure of spoken proficiency levels. Statistically, there is a high correlation between speaking proficiency and the number of words in second language (L2) speech (Kobayashi & Abe, 2016; Kobayashi et al., 2018). In the initial stages of language acquisition, an increase in running words in a limited amount of time can be one of the best indicators of language development. In the later stages, the number of words can reflect syntactic complexity in L2 speech. In other words, we can assume that word tokens are an objective and consistent measurement of L2 speaking ability. In this study, we investigate the strength of word tokens as a measuring tool with the aim of seeing how we can use it as a valuable index.

2. Background

2.1 L2 Developmental Measures

Since the 1970s, SLA researchers have sought the best “yardstick” to measure L2 development (Larsen-Freeman, 1978). Traditionally, they have focused on the T-unit (Hunt, 1970) and the average length of error-free T-units (Larsen-Freeman & Strom, 1977) as developmental indices for L2 writing. In line with these studies, Wolfe-

Quintero et al. (1998) suggested that T-unit length, error-free T-unit length, and clause length can be considered the best measures for fluency. Since then, a number of developmental studies have investigated the dimensions of complexity, accuracy, and fluency (CAF or CALF when lexis is seen as an independent domain), to assess the quality of L2 speech and writing (Housen et al., 2012). While T-unit and CAF measures have been widely used in SLA studies, the debate about their validity and universality continues (Ortega, 2003; Norris & Ortega, 2009).

L2 developmental studies have greatly benefited from learner corpus research (LCR). The availability of learner corpora enables language researchers to empirically track the language acquisition process. Additionally, the development of natural language processing technology has made it possible to analyze a broad range of linguistic features as well as several types of language errors that occur in corpora. For example, Garner and Crossley (2018) examined the growth of n-gram use in multiple indices (frequency, association strength, proportion) in the spoken performance of L2 speakers over a period of four months; they subsequently demonstrated that the frequency and proportion of bigrams were strongly related to the learners' proficiency levels. Kyle and Crossley (2018) compared traditional indices of syntactic complexity (e.g., mean length of T-units), fine-grained indices of clausal complexity, and fine-grained indices of phrasal complexity, and showed that fine-grained indices of phrasal complexity were better predictors of L2 writing quality than the other two indices. Díez-Bedmar and Pérez-Paredes (2020) analyzed noun phrase syntactic complexity in L2 writing and suggested that *nouns and modifiers* and *determiner + multiple premodification + head* can be used as indices of syntactic complexity. Meunier and Littré (2013) tracked learners' longitudinal progress in the acquisition of the English tense and aspect system and reported that tense and aspect errors decrease over time. Thewissen (2013) investigated more than 40 types of errors in essays written by learners with different proficiency levels and indicated that there is a difference in the error patterns between B1 and B2 levels of the Common European Framework of Reference for Languages (CEFR). Other learner corpus studies have explored various developmental indices, such as pragmalinguistic features (Miura, 2020) and metadiscourse markers (Kobayashi, 2017), from the perspectives of pragmatics and discourse analysis respectively. However, most studies on developmental indicators have focused on L2 writing, with fewer based on L2 speaking.

2.2 Biber's Linguistic Features

The linguistic features used by Biber (1988) aid in providing a comprehensive description of L2 speaking development. His selected set of linguistic features is broadly used in corpus-based studies to explore various types of linguistic variation (Conrad & Biber, 2001; Frignal, 2013; Sardinha & Pinto, 2014, 2019). This trend can be applied to learner corpus studies to help in identifying linguistic features that can predict the development of learners' speech (Abe, 2014), and automatically assess L2 spoken performance (Kobayashi & Abe, 2016). In this study, 67 linguistic features from Biber (1988) were used as variables to predict the increase of words in L2 spoken performance. As Table 1 shows, these features can be classified into 16 major grammatical categories: (a) tense and aspect markers, (b) place and time adverbials, (c) pronouns and pro-verbs, (d) questions, (e) nominal forms, (f) passives, (g) stative forms, (h) subordination, (i) prepositional phrases, adjectives, and adverbs, (j) lexical specificity, (k) lexical classes, (l) modals, (m) specialized verb classes, (n) reduced forms and dispreferred structures, (o) coordination, and (p) negation. Given the diversity of linguistic features to be considered, high-dimensional statistical methods that can handle a large number of variables and identify a smaller number of important variables among many features are needed.

Table 1. The 67 linguistic features from Biber (1988)

A. Tense and aspect markers
1. past tense (VBD), 2. perfect aspect (PEAS), 3. present tense (VPRT)
B. Place and time adverbials
4. place adverbials (PLACE), 5. time adverbials (TIME)
C. Pronouns and pro-verbs
6. first person pronouns (FPP1), 7. second person pronouns (SPP2), 8. third person personal pronouns (excluding *it*) (TPP3), 9. pronoun *it* (PIT), 10. demonstrative pronouns (DEMP), 11. indefinite pronouns (INPR), 12. pro-verb *do* (PROD)
D. Questions
13. direct WH-questions (WHQU)
E. Nominal forms
14. nominalizations (ending in *-tion, -ment, -ness, -ity*) (NOMZ), 15. gerunds (GER), 16. total other nouns (NN)
F. Passives
17. agentless passives (PASS), 18. *by*-passives (BYPA)
G. Stative forms
19. *be* as main verb (BEMA), 20. existential *there* (EX)
H. Subordination features

21. *that* verb complements (THVC), 22. *that* adjective complements (THAC), 23. WH clauses (WHCL), 24. infinitives (*to*-clause) (TO), 25. present participial clauses (PRESP), 26. past participial clauses (PASTP), 27. past participial WHIZ deletion relatives (WZPAST), 28. present participial WHIZ deletion relatives (WZPRES), 29. *that* relative clauses on subject position (TSUB), 30. *that* relative clauses on object position (TOBJ), 31. WH relatives on subject position (WHSUB), 32. WH relatives on object position (WHOBJ), 33. pied-piping relative clauses (PIRE), 34. sentence relatives (SERE), 35. causative adverbial subordinators (*because*) (CAUS), 36. concessive adverbial subordinators (*although, though*) (CONC), 37. conditional adverbial subordinators (*if, unless*) (COND), 38. other adverbial subordinators (OSUB)

I. Prepositional phrases, adjectives, and adverbs

39. total prepositional phrases (PIN), 40. attributive adjectives (JJ), 41. predicative adjectives (PRED), 42. total adverbs (RB)

J. Lexical specificity

43. type/token ratio (TTR), 44. mean word length (AWL)

K. Lexical classes

45. conjuncts (CONJ), 46. downtoners (DWNT), 47. hedges (HDG), 48. amplifiers (AMP), 49. emphatics (EMPH), 50. discourse particles (DPAR), 51. demonstratives (DEMO)

L. Modals

52. possibility modals (POMD), 53. necessity modals (NEMD), 54. predictive modals (PRMD)

M. Specialized verb classes

55. public verbs (PUBV), 56. private verbs (PRIV), 57. suasive verbs (SUAV), 58. *seem* and *appear* (SMP)

N. Reduced forms and dispreferred structures

59. contractions (CONT), 60. subordinator *that* deletion (THATD), 61. stranded prepositions (STPR), 62. split infinitives (SPIN), 63. split auxiliaries (SPAU)

O. Coordination

64. phrasal coordination (PHC), 65. independent clause coordination (ANDC)

P. Negation

66. syntactic negation (SYNE), 67. analytic negation (XX0)

Note. The abbreviations given in parentheses are the tags used in the Multidimensional Analysis Tagger (Nini, 2019).

2.3 Multifactorial Regression Analysis

A new methodological trend in LCR is multifactorial regression analysis (Gries, 2015; Gries & Deshors, 2014, 2021; Gries & Wulff, 2013; Wulff & Gries, 2015, 2019, 2021). In this statistical method, multiple variables (e.g., linguistic features, language errors) can be used to determine the behavior of a response (e.g., proficiency levels, word tokens). Moreover, it can evaluate the strength of association between the predictors and response in the context of statistical significance tests (e.g., *t*-test, Wald test). Thus, it allows us to simultaneously assess the multiple factors involved in language development without repeating mono-factorial tests. Multifactorial regression

analysis can be broadly divided into linear and nonlinear models, depending on the types of fitting methods. Linear models presuppose a linear relationship between predictor and response variables, while nonlinear models formulate various nonlinear relationships between predictor and response variables in cases where a linear relationship cannot be assumed. While linear models have one basic form (i.e., *response* = *constant* + *parameter* * *predictor* + ... + *parameter* * *predictor*), nonlinear models can take many different forms. In the SLA context, the language development process is not linear (Murakami, 2016). Specifically, in U-shaped development, the learners' accuracy is high in the beginning, but it drops temporarily before increasing again. In addition, in power-law development, the decrement in error becomes gradually smaller as the learner's proficiency increases. With the awareness of the nonlinearity in SLA, Murakami (2016) applied generalized additive mixed models to investigate the nonlinear patterns of the L2 accuracy development in English grammatical morphemes. Verspoor et al. (2021) also utilized generalized additive models to examine the nonlinear development in the mean length of T-units and the Guiraud index.

Random forest (Breiman, 2001) is one of the most powerful multifactorial techniques for analyzing such nonlinear developmental patterns. The method is an ensemble learning technique that operates by constructing a large collection of regression trees. The regression tree model is a nonlinear regression technique that visualizes a sequence of data classification in the form of a flowchart-like diagram (Breiman et al., 1984). In the random forest model, the ensemble of regression trees (the forest) is generated using the ensemble learning technique, to yield better predictive performance than can possibly be obtained from any of the constituent tree models. The bagging ensemble learning algorithm (Breiman, 1994) is widely used to synthesize multiple tree models. It generates a number of datasets using a bootstrap sampling technique, and then constructs multiple regression models based on each bootstrap sample. Following these steps, the random forest model calculates the average of the predictions of every single regression tree to make a final prediction. By combining regression tree and bagging ensemble learning techniques, the random forest model generally achieves higher levels of predictions than other machine learning techniques (Chen et al., 2020). This model can also handle thousands of predictor variables in a statistically efficient manner (e.g., bootstrap sampling, feature

sampling), as it is more robust to multicollinearity than linear regression and several other regression models. Moreover, this model can compute variable importance scores to measure the impact of each predictor variable on the alternation, given all other predictors. Because of these advantages, random forest is regarded as a useful tool for the identification of L2 developmental indices.

The use of random forest models has been increasing in the field of corpus linguistics. For instance, Tono (2013) applied this technique to investigate several types of language errors that occur in L2 writing and found that the omission errors of *have* and *want* are the two most important predictors of English proficiency levels. Additionally, Kobayashi and Abe (2016) predicted the quality of L2 speech using random forest and showed that word tokens and types are the best predictors of speaking proficiency. In addition to these learner corpus studies, random forest has been utilized for studies in language usage, such as verb-object-particle vs. verb-particle-object alternation (Deshors, 2019), and the choice between the progressive and simple aspects (Hundt et al., 2020).

3. Purpose of the Study

As mentioned above, word tokens can be an alternative measure of L2 speaking proficiency from a statistical perspective. Therefore, adequate predictors of word tokens in learners' spoken performance can help SLA researchers in understanding proficiency. Against this background, the present study aimed to investigate the mechanism of the increase in word tokens in L2 speaking. The research questions (RQ) that drive this article are as follows:

RQ 1: How highly correlated is the number of words with L2 speaking proficiency?

RQ 2: Which linguistic features can contribute to an increase in the number of words in L2 speech?

By pursuing RQ 1, this study validates the effectiveness of the number of words as developmental measure for L2 speaking. In addition, the answer to RQ 2 can contribute to L2 speaking assessment including automated speech scoring.

4. Methods

4.1 Corpus

The spoken data utilized in this study were extracted from the Longitudinal Corpus of L2 Spoken English (LOCSE; Abe & Kondo, 2019). LOCSE was designed to describe L2 developmental patterns, not only at the group level, but also at the individual level. The speech samples were collected from upper-secondary school students. They were public senior high school students aged 15 years at the beginning of data collection. The students spoke Japanese as their mother tongue and had no long-term experience in English-speaking countries. Additionally, they were studying the target language under a similar learning setting and had limited opportunities to speak the target language inside and outside the classroom.

The students were asked to take a monologue speaking test, the Telephone Standard Speaking Test (TSST), which consists of multiple tasks (e.g., description, comparison, reasoning). Their utterances were compiled to create the corpus data. The automated telephone-based English-speaking test consists of ten recorded questions, and test-takers were required to respond to each question in 45 seconds without any planning time or use of reference material. Three certified raters gave a holistic score to each speech sample, based on various criteria such as function-based ability, sentence structure, accuracy, and content. The test scores were divided into nine levels, ranging from level 1 (novice) to level 9 (advanced).

The speech samples collected in the test were transcribed by four trained transcribers using automated speech recognition technology (IBM Watson Speech-to-Text). For the transcription, the XML format was chosen for the interchangeability of the resource, and the annotation schema of Izumi et al. (2004) was used for comparison with other learner corpora (e.g., the NICT-JLE Corpus, Konan-JIEM Learner Corpus, KIT Speaking Test Corpus).

This study analyzed speech samples from 104 students (47 boys and 57 girls) who had taken all eight speaking tests, making a total of 832 samples. However, this study did not make use of longitudinal information of this learner corpus. Table 2 summarizes the numbers and percentages of speech samples and words for each speaking proficiency level. As the table indicates, all learners were classified into TSST levels 2–7, which correspond to the CEFR levels A1–B1. As mentioned, this study used

the number of words as a criterion for assessing language development instead of proficiency levels (for an approach that uses proficiency as a criterion, refer to Kobayashi et al., 2018; for the longitudinal analysis of the LOCSE data, refer to Abe & Kondo, 2019).

Table 2. The numbers of speech samples and words in the LOCSE

TSST level	Number of speech samples		Number of words	
2	8	(0.96%)	762	(0.21%)
3	204	(24.52%)	63,313	(17.07%)
4	468	(56.25%)	207,654	(55.99%)
5	122	(14.66%)	75,836	(20.45%)
6	27	(3.25%)	20,835	(5.62%)
7	3	(0.36%)	2,485	(0.67%)
Total	832	(100.00%)	370,885	(100.00%)

4.2 Text Preprocessing

Before analyzing the transcribed speech samples, text preprocessing was conducted. Specifically, (a) fillers (e.g., *ah, eh, umm*), (b) Japanese words excluding proper nouns (e.g., *desu, kore, nandaro*), (c) words that the transcribers could not easily identify, (d) non-verbal phenomena (e.g., cough, laughter, sigh), (e) repetitions (e.g., *he he*), and (f) self-corrections of two words or less (e.g., *I I don't like cats but I like I like dogs*) were deleted. By removing these utterances, we can count learners' pruned tokens without dysfluency markers. Furthermore, this preprocessing can increase the accuracy of natural language processing, including part-of-speech tagging and syntactic parsing.

4.3 Data Analysis

This study counted the frequencies of Biber's (1988) linguistic features using the Multidimensional Analysis Tagger (Nini, 2019) and used the frequencies for correlation analysis and random forest regression analysis. All statistical analyses in this study were conducted using R, a free software environment for statistical computing and graphics (R Core Team, 2020). The *randomForest* package (Liaw & Wiener, 2002) was used to perform the analysis. For other R techniques, including correlation analysis and data visualization, this study mainly referred to Baayen (2008) and Levshina (2015).

5. Results

5.1 Correlation Analysis

The current study begins by investigating the correlation between learners' TSST levels and word tokens using Spearman's rank correlation coefficient. As a result, speaking proficiency was found to be highly correlated with the number of words in L2 spoken performance (ρ = 0.73). This means that word tokens can function as an alternative measure for TSST levels.

As a next step, we checked the correlations among Biber's linguistic features using Pearson's product-moment correlation coefficient. Table 3 lists the 20 pairs with the highest correlations. As the table shows, the highest correlation pair among features is contraction (CONT) and analytic negation (XX0) (r = 0.76), followed by *be* as main verb (BEMA) and predicative adjectives (PRED) (r = 0.69), and subordinator *that* deletion (THATD) and private verbs (PRIV) (r = 0.64). According to the correlation coefficients, the mean word length (AWL) in L2 speech increased with the number of nouns (NN) (r = 0.42) and fell with the repetition of first-person pronouns (FPP1) (r = −0.35). The type/token ratio (TTR) also decreased through the frequent use of first-person pronouns (r = −0.36).

Table 3. The 20 highest correlation pairs of linguistic features

Rank	Variable 1	Variable 2	r	Rank	Variable 1	Variable 2	r
1	CONT	XX0	0.76	11	BEMA	PIT	0.34
2	BEMA	PRED	0.69	12	PRED	PIT	0.32
3	THATD	PRIV	0.64	13	VPRT	AMP	0.31
4	BEMA	VPRT	0.45	14	PRED	AMP	0.31
5	NN	AWL	0.42	15	PHC	NN	0.31
6	VPRT	VBD	−0.36	16	EMPH	AMP	−0.31
7	FPP1	TTR	−0.36	17	FPP1	EMPH	−0.31
8	VPRT	PIN	−0.36	18	PIN	EMPH	0.31
9	FPP1	AWL	−0.35	19	JJ	AWL	0.29
10	VPRT	PRED	0.35	20	RB	NN	−0.29

5.2 Random Forest Regression Analysis

Given the high correlation of several of the pairs shown in Table 3, this study performed a random forest regression analysis that is relatively robust to

multicollinearity in the prediction. The random forest model used Biber's linguistic features as predictor variables and word tokens as the response variable. While running the statistical algorithm, the hyperparameters of the model (e.g., the number of trees and predictor variables randomly sampled as candidates for each tree) were tuned through the *tuneRF* function of the *randomForest* package. As a result of the tuning, the model generated 500 trees using 22 variables each and explained 58.73% of the total variance of the data.

The random forest model also estimated the importance of predictor variables using the increased node impurity index (IncNodePurity). Figure 1 shows the top 30 important linguistic features in the prediction of learners' word tokens. Variables that could predict the number of words in L2 speech were, in order of strength, frequency of causative adverbial subordinators (CAUS), independent clause coordination (ANDC), emphatics (EMPH), nouns (NN), prepositional phrases (PIN), and present tense (VPRT).

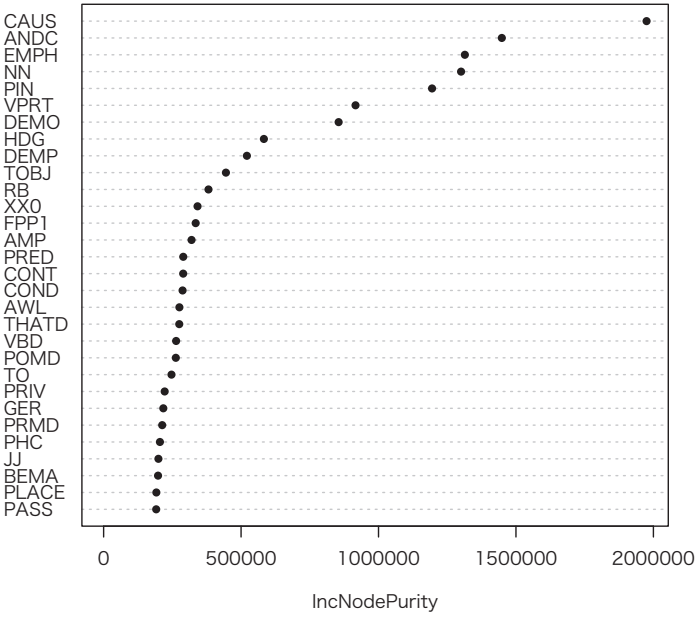

Figure 1. Variable importance plot of the top 30 linguistic features

Although there is no theoretical threshold that can be used to discriminate between important and unimportant variables, this study focuses on the top six linguistic features for detailed analysis. Figure 2 presents partial dependence plots that show how these six features affect the prediction of word tokens by marginalizing (averaging) out the effects of other features. By checking partial dependence plots, in addition to the variable importance plot, we can investigate the predictor variables while controlling for the effects of other variables (Hastie et al., 2009). The horizontal axes in the plots indicate the relative frequency of a particular linguistic feature (per 100 words), while the vertical axes indicate the number of tokens. As these plots illustrate, CAUS, NN, and VPRT are negatively related to word tokens, while EMPH and PIN are positively related. Additionally, the relative frequency of ANDC increases rapidly to around 0.1 and then decreases rapidly before it stabilizes, and it can discriminate learners in a specific range of word tokens. Interpreting the pattern in ANDC is more difficult than the patterns in the other items, but this is not because of a problem with our data. When predicting some natural phenomena, there are not many predictor variables that have values directly or inversely proportional to the values of the response variable. In the case of L2 assessment, there are some predictor variables that discriminate between learners who are above a certain level and those who are

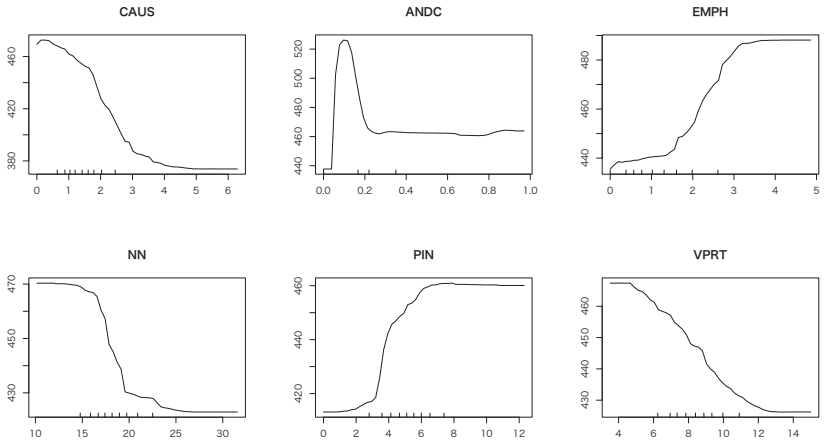

Figure 2. Partial dependence plots of the top six linguistic features

below a certain level. There are also some predictor variables that discriminate only between certain levels, such as the ANDC in this study. In other words, the random forest model provides highly accurate predictions by integrating the information held by these variables.

6. Discussion

Following the calculation of variable importance and partial dependence scores, this section explores the six important features that can predict the number of words in learners' utterances. The validity of these developmental indices will be further supported by checking concordance lines. First, the decrease in CAUS is attributed to the diversification of conjunctions that learners can use. As proficiency increases, learners can progressively construct speech without relying on the subordinating conjunction *because*. In other words, they move from the stage of "giving a reason" (e.g., *I like rainy day because rainy day is cool*) to the stage of "stating a result" (e.g., *Rainy day is cool, so I like rainy day*). Second, novice learners use ANDC with high frequency (e.g., *My mother is very careful woman, and she can find a lot of my mistakes, and she always advise me to improve my something, so I'm very I owe to her to improve my power of academic skills, and I'm very grateful for her*). After this stage, they will be able to use concessive adverbial subordinators (CONC), conditional adverbial subordinators (COND), and other adverbial subordinators (OSUB). Third, the increase in EMPH (e.g., *really, just, most, more*) allows advanced learners to express the degree of certainty in propositions more clearly. This rhetorical device can be a developmental index for both the dialogue speaking test (the Standard Speaking Test; Kobayashi & Abe, 2016) and the monologue test used in this study (the TSST). Fourth, the high frequency of NN is a prominent feature among novice learners (e.g., *I study ... five subject ... English ... Japanese, Math, and ... Science, and ... also ...*). They heavily depend on nouns in the initial stage of learning, but gradually become able to employ a variety of word types (Tono, 2000). Fifth, the increase in PIN results from the development of noun phrase structure. Additionally, prepositions become more frequent owing to the acquisition of group prepositions (e.g., *a lot of, because of*). Lastly, the decrease in VPRT use is a consequence of the increase of other tense use (e.g., *enjoyed, experienced, happened, tried*). As Table 3 shows, the frequency of the present tense is

negatively correlated with that of the past tense ($r = -0.36$).

7. Conclusion

This study aimed to explore the mechanism underlying the increase in number of words in L2 speaking. The results show that word tokens can function as an L2 developmental measure that highly correlates with speaking proficiency ($\rho = 0.73$). The results also suggest (a) causative adverbial subordinators, (b) independent clause coordination, (c) emphatics, (d) nouns, (e) prepositional phrases, and (f) present tense from Biber's linguistic features best predict the language development. These six key features can be robust measures of L2 spoken development, as they are frequently used in almost all speaking contexts. In addition, this study scrutinized the effects of these features on the increase in word tokens, by checking partial dependence plots. However, this study has some limitations. First, the frequencies of linguistic features may be affected by the tasks and topics of the TSST. Thus, we should investigate the effects of tasks and topics on learners' performance using multilevel analysis in the future. Second, the target learners were limited to novice and intermediate Japanese learners of English. It would be desirable to investigate a wider range of L1 backgrounds and proficiency levels to gain a broader understanding of the increase in word tokens in L2 speech. Third, other linguistic features can be useful for modeling the development of L2 spoken English. In particular, lexical and grammatical errors highlight language development from different angles than Biber's framework (Abe, 2007). Finally, because random forest is based on ensemble learning, a full interpretation of the results is difficult. One possible solution to this problem is to use global surrogate models that are trained to approximate the predictions of random forest models (Gries, 2020). Despite these limitations, the findings of the current study offer valuable new insights into the mechanism of the number of words in learners' speech as well as enhancing the methodology of L2 developmental studies.

Acknowledgements

This research has been partly funded by Grants-in-Aid for Scientific Research Grant Numbers 18K00849, 20K00813, and 21K00660, which we gratefully acknowledge. We are also very grateful to all the participants and researchers for their

help in the compilation of the LOCSE.

References

Abe, M. (2007) "A Corpus-based Investigation of Errors across Proficiency Levels in L2 Spoken Production." *JACET Journal 44*: 1–14.

Abe, M. (2014) "Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech." *Journal of Applied Language Studies 8*, 3: 85–96.

Abe, M., and Y. Kondo (2019) "Constructing a Longitudinal Learner Corpus to Track L2 Spoken English." *Journal of Modern Languages 29*: 23–44.

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Breiman, L. (1994) "Bagging Predictors." *Machine Learning 24*, 2: 123–140.

Breiman, L. (2001) "Random Forests." *Machine Learning 45*, 1: 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984) *Classification and Regression Trees*. Boca Raton: Chapman and Hall.

Chen, R.-C., C. Dewi, S.-W. Huang, and R. E. Caraka (2020) "Selecting Critical Features for Data Classification Based on Machine Learning Methods." *Journal of Big Data 7*, 52: 1–26.

Conrad, S., and D. Biber (Eds.) (2001) *Variation in English: Multi-dimensional Studies*. London: Longman.

Crossley, S. A., and D. S. McNamara (2012) "Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication." *Journal of Research in Reading 35*, 2: 115–135.

Deshors, S. C. (2019) "English as a Lingua Franca: A Random Forests Approach to Particle Placement in Multi-speaker Interactions." *International Journal of Applied Linguistics 30*, 2: 214–231.

Díez-Bedmar, M. B., and P. Pérez-Paredes (2020) "Noun Phrase Complexity in Young Spanish EFL Learners' Writing: Complementing Syntactic Complexity Indices with Corpus-driven Analyses." *International Journal of Corpus Linguistics 25*, 1: 4–35.

Frignal, E. (2013) "Twenty-five Years of Biber's Multi-dimensional Analysis: Introduction to the Special Issue and an Interview with Douglas Biber." *Corpora 8*, 2: 137–152.

Garner, J., and S. Crossley (2018) "A Latent Curve Model Approach to Studying L2 N-gram Development." *The Modern Language Journal 102*, 3: 494–511.

Gries, S. Th. (2015) "Statistics for Learner Corpus Research." In Granger, S., G. Gilquin and F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 159–182.

Gries, S. Th. (2020) "On Classification Trees and Random Forests in Corpus Linguistics:

Some Words of Caution and Suggestions for Improvement." *Corpus Linguistics and Linguistic Theory 16*, 3: 617–647.

Gries, S. Th., and S. C. Deshors (2014) "Using Regressions to Explore Deviations between Corpus Data and a Standard/target: Two Suggestions." *Corpora 9*, 1: 109–136.

Gries, S. Th., and S. C. Deshors (2021) "Statistical Analyses of Learner Corpus Data." In Tracy-Ventura, N., and M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, pp. 119–132.

Gries, S. Th., and S. Wulff (2013) "The Genitive Alternation in Chinese and German ESL Learners: Towards a Multifactorial Notion of Context in Learner Corpus Research." *International Journal of Corpus Linguistics 18*, 3: 327–356.

Hastie, T., R. Tibshirani, and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. New York: Springer.

Housen, A., F. Kuiken, and I. Vedder (2012) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins.

Hundt, M., P. Rautionaho, and C. Strobl (2020) "Progressive or Simple? A Corpus-based Study of Aspect in World Englishes." *Corpora 15*, 1: 77–106.

Hunt, K. W. (1970) "Do Sentences in the Second Language Grow Like Those in the First?" *TESOL Quarterly 4*, 3: 195–202.

Izumi, E., K. Uchimoto, and H. Isahara (2004) *A Speaking Corpus of 1,200 Japanese Learners of English.* Tokyo: ALC Press.

Kobayashi, Y. (2017) "Developmental Patterns of Metadiscourse in Second Language Writing." *Journal of Pan-Pacific Association of Applied Linguistics 21*, 2: 41–54.

Kobayashi, Y., and M. Abe (2016) "Automated Scoring of L2 Spoken English with Random Forests." *Journal of Pan-Pacific Association of Applied Linguistics 20*, 1: 55–73.

Kobayashi, Y., Y. Kondo, and M. Abe (2018) "Predicting EFL Learners' Oral Proficiency Levels in Monologue Tasks." In Tono, Y., and H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistic Conference*, pp. 231–236.

Kyle, K., and S. A. Crossley (2018) "Measuring Syntactic Complexity in L2 Writing Using Fine-grained Clausal and Phrasal Indices." *The Modern Language Journal 102*, 2: 333–349.

Kyle, K., S. Crossley, and M. Verspoor (2021) "Measuring Longitudinal Writing Development Using Indices of Syntactic Complexity and Sophistication." *Studies in Second Language Acquisition 43*, 4: 781–812.

Larsen-Freeman, D. (1978) "An ESL Index of Development." *TESOL Quarterly 12*, 4: 439–448.

Larsen-Freeman, D., and V. Strom (1977) "The Construction of a Second Language Acquisition Index of Development." *Language Learning 27*, 1: 123–134.

Levshina, N. (2015) *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Liaw, A., and M. Wiener (2002) "Classification and Regression by randomForest." *R News 2*, 3: 18–22.

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.

Meunier, F., and D. Littré (2013) "Tracking Learners' Progress: Adopting a Dual 'Corpus cum Experimental Data' Approach." *The Modern Language Journal 97*, S1: 61–76.

Miura, A. (2020) "Criterial Pragmalinguistic Features of Requestive Speech Acts Produced by Japanese Learners of English." *Learner Corpus Studies in Asia and the World 4*: 1–23.

Murakami, A. (2016) "Modeling Systematicity and Individuality in Nonlinear Second Language Development: The Case of English Grammatical Morphemes." *Language Learning 66*, 4: 834–871.

Nini, A. (2019) "The Multi-Dimensional Analysis Tagger." In Sardinha, T. B., and M. V. Pinto (Eds.), *Multi-dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, pp. 67–94.

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

Ortega, L. (2003) "Syntactic Complexity Measures and Their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing." *Applied Linguistics 24*, 4: 492–518.

R Core Team (2020) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/

Sardinha, T. B., and M. V. Pinto (Eds.) (2014) *Multi-dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins.

Sardinha, T. B., and M. V. Pinto (Eds.) (2019) *Multi-dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic.

Thewissen, J. (2013) "Capturing L2 Accuracy Developmental Patterns: Insights from an Error-tagged EFL Learner Corpus." *The Modern Language Journal 97*, S1: 77–101.

Tono, Y. (2000) "A Corpus-based Analysis of Interlanguage Development: Analysing Part-of-speech Tag Sequences of EFL Learner Corpora." In Lewandowska-Tomaszczyk, B., and J. P. Melia (Eds.), *PALC'99: Practical Applications in Language Corpora*. Frankfurt am Main: Peter Lang, pp. 323–340.

Tono, Y. (2013) "Criterial Feature Extraction Using Parallel Learner Corpora and Machine Learning." In Díaz-Negrillo, A., N. Ballier, and P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, pp. 169–204.

Tracy-Ventura, N., A. Huensch, and R. Mitchell (2021) "Understanding the Long-term Evolution of L2 Lexical Diversity: The Contribution of a Longitudinal Learner Corpus." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, pp. 148–171.

Verspoor, M., W. Lowie, and M. Wieling (2021) "L2 Developmental Measures from a Dynamic Perspective." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research*

Meets Second Language Acquisition. Cambridge: Cambridge University Press, pp. 172–190.

Vyatkina, N. (2013) "Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus." *The Modern Language Journal 97*, S1: 11–30.

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

Wulff, S., and S. Th. Gries (2015) "Prenominal Adjective Order Preferences in Chinese and German L2 English: A Multifactorial Corpus Study." *Linguistic Approaches to Bilingualism 5*, 1: 122–150.

Wulff, S., and S. Th. Gries (2019) "Particle Placement in Learner Language." *Language Learning 69*, 4: 873–910.

Wulff, S., and S. Th. Gries (2021) "Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, pp. 191–213.

（小林雄一郎　日本大学　Email: kobayashi.yuuichirou@nihon-u.ac.jp）
（阿部真理子　中央大学　Email: abe.127@g.chuo-u.ac.jp）
（近藤　悠介　早稲田大学　Email: yusukekondo@waseda.jp）