Vocabulary Learning and Instruction
Volume 11, Issue 2, December 2022
http://vli-journal.org

Modeling Lexical and Phraseological Sophistication in Oral Proficiency Interviews: A Conceptual Replication

Masaki Eguchi
University of Oregon

Abstract

Building on previous studies investigating the multidimensional nature of lexical use in task-based L2 performance, this study clarified the roles that the distinct lexical features play in predicting vocabulary proficiency in a corpus of L2 Oral Proficiency Interviews (OPI). A total of 85 OPI samples were rated by three separate raters based on a Common European Frame of Reference (CEFR) based rubric in terms of their linguistic range. The interview transcription was analyzed for 56 lexical and phraseological indices using modern natural language processing tools. The result of an exploratory factor analysis (EFA) revealed that the 56 indices tapped into 10 distinct factors of lexical use in OPI: three factors related to content words, three related to n-grams, three lexical collocation factors, and one function-word factor. A subsequent Bayesian mixed-effect ordinal regression indicated that six out of the 10 factors meaningfully predicted the CEFR levels on Range with reasonable accuracy (quadratic kappa coefficient = .81 with the human rating). The result highlights the distinct roles that multiple content-word, collocation, and function-word factors play in characterizing the linguistic range in a CEFR-based assessment of OPI. The implication for the assessment of lexical richness, as well as future directions of this research domain, are discussed.

Keywords: Lexical sophistication, Oral Proficiency Interview, Exploratory Factor Analysis, Bayesian mixed-effect ordinal regression

1 Introduction

A great deal of research on vocabulary acquisition and assessment has centered on conceptualizing dimensions of lexical use in task-based performance and capturing developmental changes in lexical use (Crossley et al., 2011; Laufer & Nation, 1995; Meara & Bell, 2001). There is a consensus that lexical complexity (also known as richness) can be conceptualized in terms of at least three distinct constructs: density, diversity, and sophistication (Bulté & Housen, 2012; Lu, 2012; Read, 2000). Lexical density refers to the ratio of content words in the running text. Although not used frequently, lexical density is generally reported to be higher in written texts than in spoken texts (e.g., Halliday, 1985). Lexical diversity essentially concerns the variety of lexical items used during the performance

Please cite this article as: Eguchi, M. (2022). Modeling lexical and phraseological sophistication in oral proficiency interviews: A conceptual replication. *Vocabulary Learning and Instruction, 11*(2), 1–16. https://doi.org/10.7820/vli.v11.2.Eguchi

(for more precise definitions, see Jarvis, 2013). Diversity, or more specifically lexical variation, has been traditionally operationalized using variants of the Type-Token Ratio (for a recent study, see Zenker & Kyle, 2021). Lexical sophistication attempts to capture "advanced" vocabulary use during the performance (Laufer & Nation, 1995). Recent advances in the measurement frameworks as well as operationalizations (e.g., Kyle et al., 2018; McCarthy & Jarvis, 2010) have helped uncover important dimensions in vocabulary use relating to proficiency and development (Eguchi & Kyle, 2020; Kim et al., 2018; Kyle et al., 2021). This study focuses on lexical sophistication as one important dimension of lexical use in task-based performance (Bulté & Housen, 2012; Read, 2000).

Recently, research on the relationships between lexical sophistication and second language (L2) proficiency and development benefits from a multidimensional approach to this construct (Eguchi & Kyle, 2020; Kim et al., 2018). While researchers traditionally operationalized "advanced" lexical use by counting lexical items above a certain frequency threshold (e.g., above 2,000-word level; Laufer & Nation, 1995), approaches other than the use of frequency information have been proposed as measures of this construct. Such new operationalized constructs of lexical sophistication include but are not limited to concreteness (i.e., how concrete the referent of the word is, table vs. idea; Salsbury et al., 2011), hypernymy (how specific is the denotation of the word in the word network; dachshund vs. dog; Crossley et al., 2009), contextual distinctiveness (how distinct is the context in which the word is used in a large corpus; sonic vs. chance; Berger et al., 2017), and characteristics of multiword units (e.g., Bestgen & Granger, 2014; Kyle et al., 2018).

To date, at least two studies have taken a data-driven approach to find out dimensions of lexical sophistication that determine L2 spoken and written proficiency as well as development (e.g., Kim et al., 2018). Kim et al. (2018) investigated how 100 indices of lexical sophistication (e.g., frequency, concreteness, contextual distinctiveness) calculated in the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018) characterize important dimensions of lexical use in The Yonsei English Learner Corpus (the YELC Corpus; Rhee & Jung, 2014), which includes 6,572 narrative and argumentative essays produced by Korean English a First Language (EFL) learners. Using Principal Component Analysis (PCA), they reported 12 dimensions of lexical use, 7 of which—4 content-word (CW) related dimensions and 3 multiword-related dimensions—accounted for 24.6% of the variance in the writing proficiency in the YELC corpus. As a conceptual replication in a spoken corpus, Eguchi and Kyle (2020) analyzed 1,281 Oral Proficiency Interviews (OPIs) in the National Institute of Information and Communications Technology Japanese Learner of English (NICT JLE) corpus using 91 TAALES indices. An Exploratory Factor Analysis (EFA) uncovered 10 dimensions of lexical use in the OPI, 7 of which—1 CW-related factor, 4 function-word (FW) related factors, and 2 multiword-related factors—explained approximately 57% of the variance in the OPI score. Taken together, these studies demonstrated ways in which aspects of lexical use are characterized by fine-grained lexical sophistication indices beyond the frequency of lexical items. Although these studies used a large-scale learner corpus and provided important insights into multidimensional lexical sophistication, more empirical

research is needed to confirm and refine the dimensions beyond the specific corpus. Particularly, Eguchi and Kyle (2020) extracted four factors related to FW, but they did not fully elaborate on what implication each FW factor has in terms of substantive interpretation of the model. The present study attempts to extend the findings of Eguchi and Kyle (2020) using another corpus of OPI. The study is guided by the following research questions:

1. What are the dimensions of lexical sophistication in OPIs?
2. To what extent do the dimensions of lexical sophistication explain human ratings of the OPI?

2 Method

2.1 Oral Proficiency Interview Corpus

The current study used OPI transcripts collected as a part of a large-scale project on developing an Artificial Intelligence (AI) agent conducting and assessing OPIs, the Intelligent Language Learning Assistant (InteLLA; Matsuyama, 2021, Saeki, 2021). The general procedure of the OPI in InteLLA was modelled after the ACTFL-OPI, where the interviewer adjusts the assessment task levels according to the learner's performance during the interview (Saeki et al., 2022). A total of 85 interviews were drawn from a pilot study where human experimenters controlled the agent's (non-)verbal responses from a list of pre-recorded actions. In this pilot, topics were also adaptively selected by the human experimenter, which varied in the target proficiency levels—Free time, Travel (CEFR A level), Movies, Social Networking Sites (CEFR B level), Globalization (CEFR C level) (For details of task sequence used in the OPI, see Saeki et al., 2021). In this study, given that each topic was too short to be analyzed separately, the entire OPI was used as the unit of analysis.

2.2 Range Rating

Each OPI was rated by three raters according to a CEFR-based analytic rating scale (Council of Europe, 2018), including: range, accuracy, fluency, phonology, interaction, coherence, and overall (Saeki et al., 2021). In this study, ratings on *range* were used as it is most relevant to the construct of lexical complexity. Because there were only 4 data points judged at the CEFR C2 level, C1 and C2 were combined and labelled as C-level (for the distributions of CEFR levels across three raters, see Table 1). The quadratic kappa coefficient (calculated over pairs

Table 1. Distributions of CEFR Levels Judged by Three Raters

Rater	A1	A2	B1	B2	C1	C2
A	9	14	33	18	10	1
B	7	18	33	17	8	2
C	2	29	33	31	8	1

of three raters) ranged from 0.76–0.85. Scores of all three raters' were used via mixed-effect modelling instead of taking averages (see Statistical Analysis).

2.3 Lexical Analysis

Lexical analyses were conducted with two tools: the TAALES version 2.2 (Kyle et al., 2018) and an in-house python script computing the measures of dependency collocation indices. TAALES is an open-source software that computes a range of lexical sophistication indices. I used TAALES for the reasons of comprehensiveness and reproducibility of the results. Based on the previous findings (e.g., Eguchi & Kyle, 2020; Kim et al., 2018; Kyle et al., 2018), I have chosen a total of 44 indices that represents 12 categories of lexical sophistication in TAALES. These are single-word indices such as frequency, range, contextual distinctiveness (Hoffman et al., 2013; McDonald & Shillcock, 2001), psycholinguistic word properties (Brysbaert et al., 2014; Coltheart, 1981), word recognition norms (Balota et al., 2007), age of acquisition (Kuperman et al., 2012), word neighborhood indices (Balota et al., 2007), and n-gram-based indices measuring frequency, range, and strengths of Association (SOA) (Kyle et al., 2018).

Notably, TAALES version 2.2 predates the development of more accurate measurement of lexical collocation using syntactic dependency parsers (e.g., Paquot, 2019). Thus, I supplemented 12 collocation measures using an in-house python program, four measures each for three dependency relations often used in previous research (Verb + Direct object, Adjectival modifier + Noun, and Adverbial modifier; Kyle & Eguchi, 2021; Paquot, 2019). Dependency collocation is advantageous in identifying collocations without relying on arbitrary window size, because the parser precisely identifies syntactically related words irrespective of how distant the word pairs are. For each of the dependency relations, I selected Mutual Information (MI; Church & Hanks, 1989), Mutual Information squared (MI2; Evert, 2005), and two versions of Delta P (head of the dependency as the cue, and dependency of the relations as the cue; see Gries, 2013). The SOA for this project was calculated with the SUBTLEXus corpus to closely match the target language use domain of the OPIs (Brysbaert & New, 2009).

2.4 Statistical Analysis

Exploratory Factor Analysis. EFA was run to identify the latent structure of the lexical variables. This analysis was done in the following steps. First, the factorability of the dataset was examined through the Kaiser-Meyer-Olkin (KMO) index. Second, an adequate number of factors was estimated using parallel analysis and other model fit indices with maximum likelihood estimation. Specifically, the empirical Bayesian Information Criterion (eBIC) was used to see the fit of the factor structure to the dataset. Once the factor structure is determined, the factor solution was identified using Oblimin rotation, which allows correlations among extracted factors. Finally, factor scores of each lexical performance were estimated using TenBerge estimation method, which preserves the overall factor correlations. The analysis was conducted through the *psych* package (Revelle, 2016) in R (R development Core Team, 2019).

Bayesian ordinal mixed-effect regression. The extracted factor scores were used as predictors for the model predicting the CEFR levels of the OPI performance. As CEFR levels are ordinal, an ordinal regression model was fit while raters were treated as a random effect (Baters & Vuorre, 2019). Specifically, a cumulative logit model was fit, which assumes that the observed outcome (i.e., CEFR levels on Range) follows a latent normal distribution underlyingly (i.e., proficiency). One potential issue of fitting a Bayesian model with a relatively small dataset is that the parameter estimation could be affected by the selection of prior distribution. For this reason, in addition to the prior predictive checks, a sensitivity analysis was conducted to see the impact of different prior selections on the results. For the final model presented below, student-*t* distribution with degrees of freedom of three, mean of zero, and standard deviation (SD) of two was used. The Bayesian analyses were conducted using the *brms* package (Bürkner, 2017). An effect size (pseudo-R^2) metric was computed using the Bayesian method introduced in Gelman et al. (2019) using the pseudo-R^2 formula proposed by McKelvey and Zavoina (1975), where $R^2 = \frac{\textit{Explained variance}}{\textit{Explained variance} + \textit{Residual variance}}$ with an assumption that the residual variance equals $\pi^2 / 3$ for models with the logistic link function. Further, the degree to which the predictor explains the CEFR Range rating was assessed via classification accuracy using the quadratic kappa coefficient. Due to the space limitation, readers are referred to available supplementary material detailing the analysis procedure at: https://osf.io/jrfa7.

3 Results

3.1 Exploratory Factor Analysis

The KMO indicated that the dataset is adequate for factor analysis (KMO = 0.60). A parallel analysis suggested a solution with seven factors, and eBIC index suggested a 10-factor solution. A close examination of the pattern matrices (after rotation) indicated that the 10-factor solution seemed to reflect the conceptual categories of indices better. Additional indices such as the Standardized Root Mean Squared Residual (SRMR) showed that the 7-factor solution might not adequately reproduce the variance-covariance matrix ($SRMR_{7\text{-factor}} = 0.081$; $SRMR_{10\text{-factor}} = 0.046$). For these reasons, the 10-factor solution was chosen, and the factor scores for each performance were estimated. The final 10-factor solution accounted for approximately 67% of the variance in the data. The inter-factor correlations suggest that some factors have moderate correlations; $r_{F1\&F2} = 0.49$; $r_{F3\&F5} = 0.42$ (for a full inter-factor correlation matrix, see online supplementary materials). Due to the space limitation, the following sections briefly summarize the identified factors (for a complete pattern matrix, see online supplementary material).

F1: Content Word (CW) Acquisition Properties. CW acquisition properties included indices related to SUBTLEXus logged frequency and range, phonological neighborhood indices, Academic Word List, Age of Acquisition/Exposure, and Word recognition norms. The higher the score on the CW acquisition factor,

the less frequent, more academic, more formally distinct, and later acquired content words the OPI contained.

F2: Function Word Properties. FW properties factor included FW-related indices related to imageability, concreteness, age of acquisition, phonological neighborhood, contextual distinctiveness, and word recognition norm. The higher the FW property score, the less imageable and concrete, later acquired, more formally distinct, more contextually distinctive function words the OPI contained.

F3: Trigram Frequency and Strengths of Association. Trigram Frequency and SOA factor included indices related to Trigram range and SOA, which particularly highlights more frequent combinations (Approximate Collexeme and T-score). The higher the score on this factor, the more frequent, higher associating trigrams the OPI response contained.

F4: Spoken Content Word Distributional Properties. Spoken CW distribution factor contained COCA spoken frequency and range, McDonald contextual distinctiveness (McDonald & Shillcock, 2001), and Bigram character frequency.

F5: Bigram Frequency and Strengths of Association. Bigram Frequency and SOA factor included range as well as four bidirectional SOA measures of this unit. As such, the higher the scores on this factor, the more commonly used, highly associated bigrams the OPI contained.

F6: Perceptually Salient Referents. Perceptually Salient Referents factor included concreteness and imageability of content words (both positively loaded). Since these indices relate to the perceptual salience of the referents of these words, the higher the factor score, the more perceptually recognizable words (concrete words referring to tangible objects) the OPI contained.

F7: Verb–Direct Object (Verb-dobj) Strengths of Association. Verb-dobj SOA factor mainly included SOA measures of Verb-dobj word pairs, with two additional indices related to contextual distinctiveness and familiarity. The negative factor loadings of the Semantic Diversity (SemD; Hoffman, 2013) and familiarity measures indicated that the higher score on this factor, the more semantically specific, fewer familiar words were used in the OPI. The positive loadings of Verb-dobj object SOAs would mean that high factor scores associated with the use of Verb-dobj object pairs, expressing the predicates and a core participant of the clause's prepositional content, more conventional in SUBTLEXus corpus. Taken together, the higher the Verb-dobj factor, the more semantically specific word choice was made, with arguably more conventional choices of verb-direct object pairs.

F8: Exclusive Trigram Use. Exclusive Trigram Use factor included strongly loaded two Trigram SOA measures (MI and MI^2) and moderately loaded function word indices that pertain to the formal aspect of the FW words. The higher the scores on this factor, the more conventional Trigrams the OPI contained.

F9: Adverbial Modifier Strengths of Association. Adverbial modifier SOA factor included four indices related to the Adverbial modifier word pairs and a Lexical Decision latency of function words. The higher the factor scores of adverb SOA, the more conventional adverbial modifier and their head was used in the OPI.

F10: Adjectival Modifier Strengths of Association. Adjectival modifier SOA factor included four SOA indices of this adjective + noun word pairs and a phonological neighborhood size of function words. Therefore, the higher the adjectival modifier SOA factor score, the more conventional adjective + noun word pairs were used in the OPI.

3.2 Bayesian Ordinal Regression

Using the 10 factors based on the EFA model, a series of ordinal mixed-effect regression models were constructed to examine the extent to which the lexical use explains the CEFR levels. Since by-rater random slopes for the lexical factors did not improve the model fit, the by-rater intercept-only model is reported as the best-fitting model. The model was assessed as converged using the R-hat values lower than 1.01 and the visual inspection of the posterior predictive distribution (Vehtari et al., 2020). The sensitivity analysis indicated that the selection of plausible alternative prior distribution did not greatly affect the substantive interpretation of the model parameters (see online supplementary material). Table 2 reports the estimated parameters of this model.

Six out of the 10 factors meaningfully predicted CEFR levels as indicated by the Credible Intervals (95% CrI; Table 2). These factors were three CW-related factors, one FW-related factor, one Bigram factor, and one dependency collocation factor (Verb–**dobj** object). Conversely, two trigram factors, adverbial modifier collocations, and adjectival modifier collocations were not meaningfully associated with the CEFR levels above and beyond other factors.

All the CW-related factors contributed to predicting the CEFR levels. For instance, the model parameter indicated that a 1.0 SD increase in CW acquisition properties was associated with a positive 1.15 SD change in the latent distribution of the CEFR levels. By exponentiating the parameter estimates (exp[1.15] = 3.158), one SD *increase* in CW acquisition factor changes the odds of being categorized in an adjacent higher level compared to the current level by 3.158 (95% $\text{CrI}_{\text{odds ratio}}$ = 2.24–4.57). Similarly, a positive effect of Spoken CW factor indicated that one SD *increase* in Spoken CW distribution resulted in 1.58 times the odds of being categorized in an adjacent category compared to the current one (95% $\text{CrI}_{\text{odds ratio}}$ = 1.19–2.09). The negative effect of Perceptually Salient Referents indicated that one unit *decrease* on this factor leads to an increased odds of being at the next category compared to the current one by 1.87 times (95% $\text{CrI}_{\text{odds ratio}}$ = 1.419–2.509; inverted before exponentiating).

Function-word properties factor, the only FW factor out of the EFA model, was positively associated with the CEFR levels. One SD *increase* in the factor would result in the changes in odds of being categorized at the next level over the current one by 4.305 times (95% $\text{CrI}_{\text{odds ratio}}$ = 3.004–6.359).

Two factors tapping into multi-word units predicted the CEFR levels. A proficient OPI would include more commonly used, strongly associated bigrams, with one SD increase on bigram frequency and SOA factor resulting in 1.99 times the odds of a performance being categorized at the next level compared to the current level (95% $\text{CrI}_{\text{odds ratio}}$ = 1.462–2.745). The negative coefficient of Verb–dobj SOA

Table 2. Summary of the Ordinal Regression Model

Predictors	LogOdds	Est.Error	CrI-Lower	CrI-Upper	R-hat	Bulk_ESS	Tail_ESS
Intercept[1]	−5.48	0.57	−6.65	−4.41	1.000	1757	2380
Intercept[2]	−2.07	0.41	−2.93	−1.29	1.000	2546	2588
Intercept[3]	−0.46	0.4	0.65	2.28	1.000	2848	2612
Intercept[4]	4.23	0.48	3.27	5.17	1.000	2821	2862
CW acquisition	1.15	0.18	0.81	1.52	1.000	3320	2738
FW property	1.46	0.19	1.1	1.85	1.000	3384	2712
Trigram Freq and SOA	−0.05	0.15	−0.35	0.25	1.000	3628	2924
Spoken CW distribution	0.46	0.14	0.18	0.74	1.000	4365	2856
Bigram Freq and SOA	0.69	0.16	0.38	1.01	1.000	3035	2845
Perceptually Salient Referents	−0.63	0.15	−0.92	−0.35	1.000	4230	3125
Dobj SOA	−0.95	0.16	−1.25	−0.64	1.000	2906	3099
Excl. Trigram	0.12	0.14	−0.16	0.4	1.000	3082	2940
Advmod SOA	−0.14	0.14	−0.42	0.13	1.000	3877	2963
Amod SOA	0.03	0.14	−0.25	0.29	1.000	3942	2942

	LogOdds	Est.Error	CrI-Lower	CrI-Upper	R-hat	Bulk_ESS	Tail_ESS
SD(Intercept)	0.48	0.52	0.01	1.93	1.000	1229	1748

	Estimate	Est.Error	CrI-Lower	CrI-Upper
Bayesian pseudo-R^2 (McKelvery & Zavoina, 1975)	0.732	0.031	0.667	0.787

Note. CrI = 95% Credible Intervals. ESS = Effective sample size (in the MCMC sample); R-hat ≤ 1.01 is considered convergence of the parameter estimates (Vehtari et al., 2020).

indicates that one unit *decrease* on this factor would result in the increase of odds being categorized at a next level compared to the current one by 2.58 times (95% $\text{CrI}_{\text{odds ratio}}$ = 1.896–3.490).

Finally, the model was assessed by examining a confusion matrix as well as the quadratic kappa coefficient, which is often used to assess inter-rater agreement. The quadratic kappa on the 255 human ratings (85 speech ∗ 3 raters) was 0.81, suggesting that the lexical factors can make a satisfactory prediction of CEFR levels. Considering that the inter-rater agreement ranged 0.76–0.85, the current model achieved a human-level agreement. A visual inspection of the confusion

Table 3. Confusion Matrix for Predicted Category

Human Rating	Predicted category				
	A1	A2	B1	B2	> C
A1	**10**	8	0	0	0
A2	3	**35**	14	0	0
B1	0	14	**73**	12	0
B2	0	1	23	**23**	9
> C	0	0	1	15	**14**

Note. k = 255 (85 learners ∗ 3 raters).

matrix (Table 3) indicated that all but two predicted data points fell within the adjacent levels.

4 Discussion and Implications

Building on previous research on aspects of lexical use and the L2 spoken proficiency (Eguchi & Kyle, 2020; Kim et al., 2018), this study investigated the dimensions of lexical use in OPI including recently developed, arguably more sophisticated measures of lexical collocations. The result of EFA indicated that an optimal factor structure with 56 measures was a 10-factor solution, with three separate CW-related, one FW-related, three *n*-gram, and three dependency collocation factors. A subsequent Bayesian ordinal mixed-effect regression indicated that 6 out of the 10 factors meaningfully contributed to predicting the CEFR levels of an OPI performance. A quadratic kappa coefficient (0.81) indicated that the model was satisfactory in reproducing the human rating on the linguistic range assessed according to the CEFR scale. In what follows, three emerging issues are discussed.

4.1 Three Dimensions of CW Use

The result indicated that three CW-related factors contributed to the regression model predicting CEFR levels. A more proficient OPI was characterized by higher scores on the CW acquisition factor (F1), Spoken word distribution (F4), and lower scores on Perceptually salient referents (F6). To illustrate the patterns highlighted by CW acquisition and Spoken word distribution, Figure 1 lists randomly sampled 35 words with scores computed using the indices loaded on each factor (not weighted according to factor loading). High scores on Y-axis mean that the words tend to be learned later (e.g., *difficulties*, *preparation*, *curiosity*), reflecting high scores on F1: CW acquisition properties. Arguably, these words may allow the speaker to be more precise when they respond to the interviews. In contrast, high scores on X-axis indicate that they are more commonly used in the Spoken corpus (e.g., *came*, *seems, main, meant*), or high scores in the F4: Spoken CW distribution. These words do not seem necessarily "advanced"

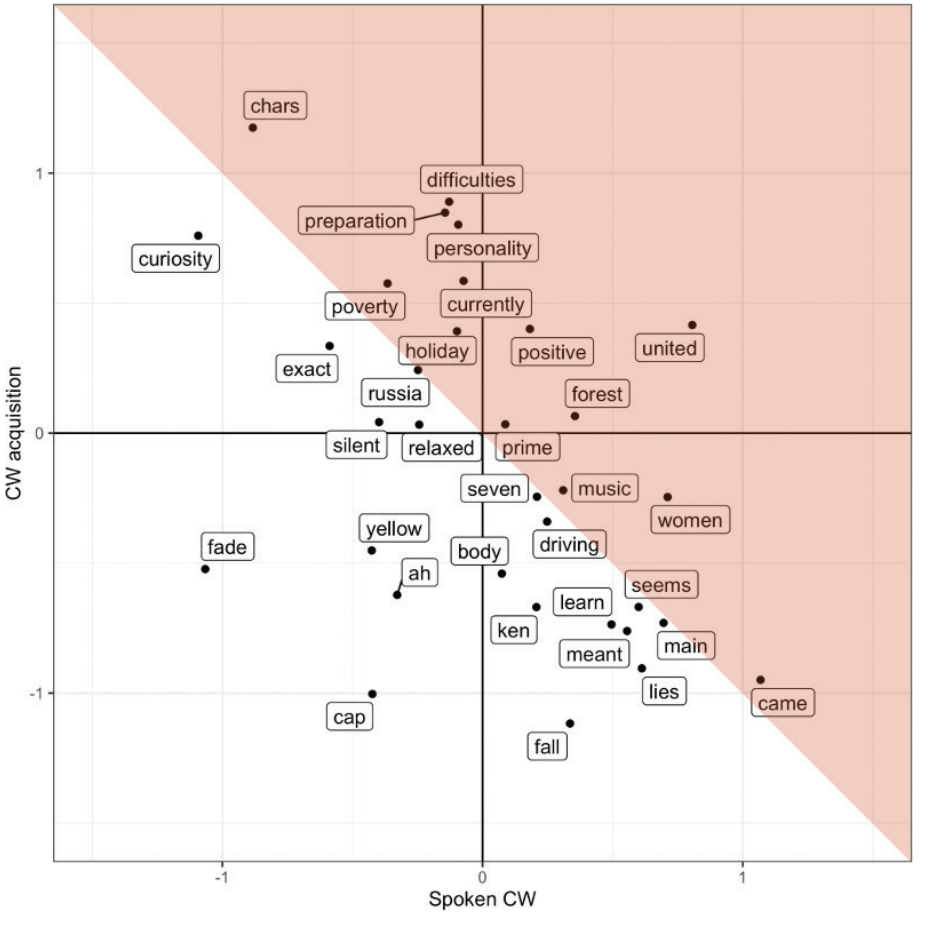

Figure 1. Example Content Words along Two Factors.

as they are frequent content words; however, they would serve critical communicative functions in the OPI discourse. This visualization appears to support the notion of the multidimensional lexical sophistication (Eguchi & Kyle, 2020; Kim et al., 2018).

4.2 Low Strengths of Association of Verb–Direct Object as Characteristics of Proficient OPIs

One finding that contrasts with previous research was the negative relationship between OPI scores and Verb–dobj SOA (Kyle & Eguchi, 2021; Paquot, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). Since the finding appears to contradict the assumption of the measure, namely, the more proficient learners should know highly conventional word combinations, a follow-up qualitative analysis was conducted. Table 4 lists 10 example collocations used in OPIs from three distinct CEFR levels. One possible reason for high SOA scores from

Table 4. Example Verb–Direct Object Collocations for Three Different Proficiency Levels

CEFR level = A		CEFR level = B		CEFR level = C	
Verb–Dobj	MI2	Verb–Dobj	MI2	Verb–Dobj	MI2
have__music	6.271	like__clothe	3.962	have__evaluation	1.87
watch__house	7.288	like__cream	8.173	get__update	4.91
buy__anything	7.614	do__job	13.743	experience__life	6.15
watch__movie	13.731	meet__you	14.495	give__feedback	6.15
meet__you	14.495	do__what	18.696	get__letter	10.18
watch__tv	16.119	like__bedroom	NA	have__opportunity	10.42
thank__you	19.097	like__summer	NA	do__that	15.81
eat__coffee	NA	go__festival	NA	expand__horizon	15.86
like__summer	NA	want__london	NA	say__university	NA
like__swimming	NA	like__culture	NA	improve__tech	NA
listen__music	NA	enjoy__culture	NA	evaluate__teacher	NA

lower CEFR levels is the use of highly formulaic language (e.g., *thank you*, *meet you*) and widely used component words occupying the direct object (e.g., *music*, *anything*). At the higher proficiency levels, the OPI tends to contain word pairs with low SOAs, (e.g., have__ evaluation), which may have lowered the average SOAs. Thus, the current result may also be due to the methodological decision because word pairs were absent from the reference corpus were not included in the denominator calculating the mean score. More research is needed to best handle word pairs that do not occur in the reference corpus.

4.3 Function Words as Characteristics of Proficiency or Task?

As found in previous studies (e.g., Eguchi & Kyle, 2020), function words were predictive of the OPI performance. At face value, this may simply indicate that highly proficient speakers are adept at using prepositions and conjunctions (see Table 5 for a randomly sampled 15 function words). A qualitative examination revealed that function words with low FW property scores included pronouns, demonstratives, and some subordinate conjunctions. On the other hand, words with higher FW property scores were those having meanings that are schematic (e.g., *among*, *through*), which are used to add oblique information (e.g., LOCATION, PATH, etc.). For one thing, this suggests that highly proficient O PIs m ay c ontain p ropositionally e laborated utterances (by way of prepositional phrases). However, this is not the only interpretation possible. That is, given that the OPI tasks were adaptively selected according to the test-takers initial turns, the tasks that are likely to elicit prepositional phrases may have been assigned to test-takers who performed reasonably well at the warm-up stage, creating a confound in the measurement. A more detailed analysis is needed to tease apart the two (potentially complementary) interpretations.

Table 5. Example Function Words with FW Property Scores

Function words	FW property
she	−0.886
like	−0.737
if	−0.287
past	−0.171
this	−0.127
when	−0.104
from	0.271
some	0.279
that	0.285
without	0.354
following	0.358
so	0.407
among	0.457
through	0.565

5 Conclusion

Before concluding the study, a few limitations and future directions are worth mentioning. First, the number of OPI interviews was limited to 85 interviews, which was quite smaller than the previous study (1,281 interviews; Eguchi & Kyle, 2020). Despite this limitation, one strength of the current study included recruiting three raters (with a completely crossed design), which allowed the estimation of the rater variabilities (see online supplementary material at w https://osf.io/jrfa7/). However, it is still ideal to increase the size of the OPI interviews to estimate the category-specific effects of lexical factors; that is, whether each lexical factor is more or less effective in differentiating the range score, and if so, at which proficiency levels. During the analysis, I indeed estimated category-specific effects for each of the 10 factors, but the model showed the tendency of overfitting, likely because of the small sample size. With increased N sizes, it would be beneficial to examine which proficiency levels each lexical use factor would have the most discriminatory power, which may in turn speak to the developmental stages of lexical performance. Second, the unit of analysis used in this study was the entire OPI, not tasks within OPIs, because of short by-task productions. The result of the study, particularly that of FW-related factors, suggests that it is worthwhile to investigate moderating effects of tasks when predicting proficiency. As such, the field would merit examining the minimum text lengths for lexical sophistication indices, which is done in lexical diversity indices (e.g., Zenker & Kyle, 2021). Although the method used to examine the stability of lexical diversity is not easily transferrable in the study of lexical sophistication, such research enables researchers to design tasks to elicit enough utterances to investigate task effects on lexical complexity.

Despite the potential limitations, the current study is generally in line with the conclusion of the previous studies (Eguchi & Kyle, 2020; Kim et al., 2018). Taken together, the multidimensional approach to lexical and phraseological sophistication would reveal salient dimensions of both context-specific and/or proficiency-distinguishing features of lexical use in task-based performance. The study further demonstrated that not only the use of content words but also that of function words and phraseological units contribute to the assessment of lexical ranges. Future studies should examine more detailed task effects and text-length effects on the measurement of lexical sophistication.

Acknowledgments

I am grateful to Yoichi Matsuyama, Shungo Suzuki, Mao Saeki, and Ryuuki Matsuura for granting me access to the OPI corpus. The author also thanks Kristopher Kyle, Joseph Vitta, and the participants at the JALT Vocab SIG Symposium 2022 for their valuable input. This paper is based on results obtained from a project, JPNP20006 ("Online Language Learning AI Assistant that Grows with People"), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Educational Technology & Society*, *20*(2), 201–212. https://eric.ed.gov/?id=EJ1137662

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Language learning & language teaching*, Vol. 32 (pp. 21–46). John Benjamins Publishing Company.

Bürkner, P.-C. (2017). brms: An R package for Bayesian Multilevel Models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Coltheart, M. (1981). The MRC Psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume with new descriptors*. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334. https://doi.org/10.1111/j.1467-9922.2009.00508.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

Evert, S. (2005). *The statistics of word cooccurrences word pairs and collocations*. Unpublished doctoral dissertation, Institut Fur Maschinelle Sprachverarbeitung Universitat Stuttgart. Retrieved from http://en.scientificcommons.org/19948039

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next ... *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Halliday, M. A. K. (1985). *Spoken and written language*. Deakin University Press.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(S1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Kim, M. M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using words, n-gram, and dependency bigram indices. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora.* Multiligual Matters.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Matsuyama, Y. (2021). Tutorial English AI: 人と共に成長するオンライン語学学習支援 AI システムの開発 [Tutorial English AI: Online Language Learning AI Assistant Growing with Humans]. 人工知能学会研究会資料 言語 音声理解と対話処理研究会 93 回 *(2021/11) [SLUD]*, 172–173.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295–322. https://doi.org/10.1177/00238309010440030101

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5–19. http://www.ameprc.mq.edu.au/__data/assets/pdf_file/0013/241411/Prospect_16,3_article_1.pdf

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

R development Core Team. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

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

Revelle, W. (2016). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. Retrieved from http://personality-project.org/r/book/

Rhee, S. C., & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Korea Contents Association*, *4*, 1019–1029. https://doi.org/10.5392/JKCA.2014.14.11.1019

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora.* (pp. 101–125). Multiligual Matters.

Saeki, M., Demkow, W., Kobayashi, T., & Matsuyama, Y. (2022). A WoZ Study for an Incremental Proficiency Scoring Interview Agent Eliciting Ratable Samples. In S. Stoyanchev, S. Ultes, & H. Li (Eds.), *Conversational AI for Natural Human-Centric Interaction* (pp. 193–201). Springer Nature. https://doi.org/10.1007/978-981-19-5538-9_13

Saeki, M., Matsuyama, Y., Kobashikawa, S., Ogawa, T., & Kobayashi, T. (2021). Analysis of Multimodal Features for Speaking Proficiency Scoring in an Interview Dialogue. *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 629–635). Shenzhen, China. https://doi.org/10.1109/SLT48900.2021.9383590

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343–360. https://doi.org/10.1177/0267658310395851

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study*. *International Journal of Learner Corpus Research*, *7*(2), 197–229. https://doi.org/10.1075/ijlcr.20015.van

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-ba1221

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505