LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 16, No: 2, July – December 2023

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

A Corpus Study of Language Simplification and Grammar in Graded Readers

Azrifah Zakaria [a,*], Willy A Renandya[b], Vahid Aryadoust[c]

[a] nie21.az2285@e.ntu.edu.sg, English Language and Literature, National Institute of Education, Nanyang Technological University, Singapore
[b] willy.renandya@nie.edu.sg, English Language and Literature, National Institute of Education, Nanyang Technological University, Singapore
[c] vahid.aryadoust@nie.edu.sg, English Language and Literature, National Institute of Education, Nanyang Technological University, Singapore
[*] Corresponding author, nie21.az2285@e.ntu.edu.sg

APA Citation:
Zakaria, A. Renandya, W. A., & Aryadoust, V. (2023). A corpus study of language simplification and grammar in graded readers. *LEARN Journal: Language Education and Acquisition Research Network, 16*(2), 130-153.

Received
24/10/2022

Received in revised form 10/02/2023

Accepted
27/03/2023

ABSTRACT

Studies on graded readers used in extensive reading have tended to focus on vocabulary. This study set out to investigate the linguistic profile of graded readers, taking into account both grammar and lexis. A corpus of 90 readers were tagged according to the variables in Biber's Multidimensional (MD) analysis, using the Multidimensional Analysis Tagger (MAT). These variables were analysed using latent class cluster analysis to determine whether the graded readers can be grouped by similarity in linguistic features. While MAT analysis surfaced more similarities than differences within the corpus, latent class clustering produced an optimal 3-class model. Post-hoc concordance analyses showed that graded readers may be categorised as having three classes of complexity: beginner, transitional, and advanced. The findings in the study suggest that selection of reading materials for extensive reading should take into consideration grammatical complexity as well as lexis. The linguistic profiles compiled in this study detail the grammatical structures and the associated lexical items within the structures that teachers may expect their students to encounter when reading graded readers. In addition, the profiles may be of benefit to teachers seeking to supplement extensive reading with form-focused instruction.

Keywords: corpus linguistics, extensive reading, graded readers, latent class analysis, multidimensional analysis

Introduction

Extensive reading (ER) has long been used in many language learning contexts, particularly in teaching English to speakers of other languages (TESOL). ER programmes serve to enrich students' classroom learning experience and their benefits are well-documented (Jeon & Day, 2016; Nakanishi, 2015). While there are many forms of ER, at its core, ER involves independent reading at levels appropriate for learners. Variously called silent sustained reading, free voluntary reading, and Drop Everything and Read (DEAR), ER has been used in many contexts, both L1 and L2. ER programmes serve to bridge the classroom and leisure, where large amounts of pleasurable reading buttress language learning. The effectiveness of ER, however, hinges on the comprehensibility of the texts – if the texts are too difficult, reading ceases to be pleasurable and learners are often put off. This, in turn, limits the amount of language exposure, and, by extension, learning.

A wide variety of texts has been used for L2 extensive reading, including language learner literature, a genre to which graded readers belong. Such texts are specifically designed to provide learners with interesting and comprehensible reading materials through language simplification. This is often accomplished by restricting vocabulary, according to word levels or some other criteria. Reading at the appropriate level results in automaticity and comprehension, contributing to a virtuous cycle where learners seek to read more, thus, enhancing their language learning (Nation & Waring, 2020). It is unsurprising, therefore, that one of the major concerns in ER research and practice is the selection of reading materials. In addition, given the frequency effects in language learning (Ellis, 2002), quantitative analyses of texts often prove fruitful in characterizing levels for reading (e.g., Crossley et al., 2011; Crossley et al., 2014).

Consequently, an understanding of the nature of the linguistic input learners are exposed to is essential before initiating an ER programme. Considerable research has gone into analysing texts used for ER but much of the extant literature is concentrated on lexis rather than grammar. What is less clear is the range of grammatical structures learners are exposed to when reading graded readers.

Study Aims and Research Questions

The purpose of this study, therefore, is to provide a more complete understanding of the linguistic features present in graded readers, accounting for both lexis and syntax. In particular, the study addresses the following questions:

RQ1. What are the linguistic features present in graded readers and how are these linguistic features distributed?

RQ2. What are the discernible characteristics of grammatical patterns in graded readers?

To answer the RQs, a corpus of graded readers was analysed using a combination of quantitative and qualitative methods. This study aims to provide some useful information towards language learning and pedagogy by deriving a list of grammatical structures present in graded readers that may be used in the classroom for explicit form-focused instruction.

Literature Review

The premise of simplified texts rests upon the central idea that words are not evenly distributed, with some words occurring very frequently and others occurring very rarely. This Zipfian distribution leads to an important area of research related to reading: the minimum threshold of lexical coverage necessary for reading comprehension. The threshold of lexical coverage refers to the minimum percentage of words in the text that a reader must know to understand it. If, the argument goes, a reader knows fewer words than this minimum threshold, they would not be able to comprehend that text. This is important for language learning and

teaching as it provides a guide for vocabulary learning. For ER, the target for learners is "only for sufficient understanding to achieve their reading purpose" (Bamford & Day, 2004, p.3). To reach 98% lexical coverage (the oft-quoted upper limit for good comprehension to occur, e.g., Laufer & Ravenhorst-Kalovski, 2010), readers would have to know 3,000-word families for graded readers (Nation, 2006).

L2 readers tend to find simplified texts easier to read, as indicated by their reading speed (Crossley et al., 2014) and learners' self-reports (Kano, 2015). Simplified texts are also better understood, although background knowledge may compensate for gaps in linguistic knowledge when students read unsimplified texts (Crossley et al., 2014). The perception of the relative ease of simplified texts may be due to the lexical differences present between the two types of texts; in Kano (2015), Youth Readers (first language (L1) children's texts) had a lower percentage of the basic 1000-word-level vocabulary that that in second language (L2) texts, even when both were categorized as having the same level of difficulty. These basic words had much higher frequencies and usage varieties (Kano, 2015). This is comparable to the findings in Crossley et al. (2012) that lexical simplification leads to higher polysemy and superordination, and lower verb hypernymy (verb specificity). Highly frequent words that are word substitutes for less frequently occurring words tend to be more polysemous and less hypernymous. Words with lower hypernymy values, greater superordination, or greater polysemy tend to be more abstract and ambiguous and may lead to more confusion for the learner (Crossley et al., 2012).

Both these studies (Kano, 2015; Crossley et al., 2012) also examined the effect of simplification on syntactic complexity. L1 texts contained more complicated sentence structures (Kano, 2015) than graded readers. Crossley et al. (2012) cautioned that syntactic simplification may result in a loss of spatial cohesion as there were fewer prepositions that relate to motion in graded readers at the beginner level, compared to more advanced graded readers. Despite this observed relationship between lexical simplification and syntactic complexity, few other studies have investigated the syntax of graded readers.

While we can examine lexis and syntax separately, the preponderance of formulaic sequences and multiword expressions have shown that lexis and grammar can be intertwined. Given the frequency of formulaic sequences, learning these non-transparent expressions may prove to be essential to comprehension (Martinez & Schmitt, 2012). Although the meanings of multiword expressions are more transparent, the frequency of co-occurrence is such that substitutions of words within multiword expressions show a lack of fluency or are deemed ungrammatical (Ellis, 2002). Promisingly, graded readers were found to contain as many or more, and similar types of, lexical bundles as the original texts although at the Common European Framework of Reference for Languages B1 level, some high frequency lexical bundles were noticeably absent (Allan, 2016). Questions remain, however, of the relationship between lexis and syntax in simplified language, especially since the latter is comparatively under-researched.

Outside of the research into this specific text type of graded readers, there has been an increasing use of automated methods in the study of linguistic variation in different genres or registers of texts. The analysis of textual difficulty and cohesion through the online programme Coh-Metrix (http://141.225.41.245/cohmetrix2017), for example, has provided much insight into linguistic features that contribute to text cohesion (e.g., Graesser et al., 2014). The Coh-Metrix indices span descriptive statistics (e.g., mean sentence length), lexical diversity and syntactic complexity. Graesser et al. (2014) asserted that "text difficulty is inherently multidimensional" (p. 212) and that no single index could measure textual difficulty. The same could be said for linguistic variation, a conclusion Biber arrived at decades prior in his seminal study (1995) *Variation across Speech and Writing*. Biber's use of factor analysis to determine patterns of co-occurrences of linguistic features hinges on a mixture of measures related to both lexis and syntax. Biber (1995) compiled a corpus of written and spoken texts and tagged the corpus for a selection of linguistic features (e.g., tense and aspect markers, place and time adverbials, pronouns and pro-verbs). Using factor analysis on the frequency counts of linguistic features to determine their patterns of co-occurrences, Biber (1995) derived dimension scores of the texts in the corpus, that, taken as a

continuum, are indicative of tendencies within a piece of text or a corpus towards certain registers. His multidimensional (MD) analysis is still in current use (e.g., Berber Sardinha et al., 2015) to explore diachronic variation (e.g., Biber & Gray, 2011), register differentiation (e.g., Berber Sardinha, & Veirano Pinto, 2019), and discipline-specific variation of academic writing (e.g., Gray, 2013), among many others. MD analysis has also been applied to study single discourse types (e.g., Friginal et al., 2013). Such studies provide an understanding on the types of linguistic features associated with particular functions and registers that can translate in practice to more specified instruction.

Beyond the dimension scores, one of the more important concepts from this seminal study is that linguistic patterns co-occur and can be "identified empirically rather than being proposed in an *a priori* functional basis" (1995, p.24). These linguistic features are first identified by word class (Biber, 1995, p. 211), then by "sequences of words" (Biber, 1995, p.212). Subsequently, Biber developed, with reference to Quirk et al. (1985), the algorithms to disambiguate grammatical categories. Taking as a point of departure this theoretical position that patterns of co-occurrence are indicative of latent structures within linguistic data, this study adopts Biber's framework to explore grammatical variation within graded readers.

Methodology

Linguistic features originally identified by Biber (1995) are tagged and analyzed using Nini's (2020) software Multidimensional Analysis Tagger (MAT). A combination of quantitative (Latent Class Cluster analysis) and qualitative analyses (of concordance lines) was used to analyse a corpus of graded readers.

Corpus of Graded Readers

The corpus comprised of 90 graded readers made available for research on Lextutor. The graded readers were originally published by Oxford University Press (OUP), Cambridge University Press (CUP) and Penguin Pearson (PP) under the imprint of Oxford Bookworms (OB), Cambridge English Readers (CER) and Penguin Readers (PR), respectively. Only fictional texts were selected for this study as the number of non-fiction texts available on the Lextutor site did not suffice for a balanced corpus. The texts were a mix of classic literary texts and modern novels, rewritten for the appropriate grade levels, as well as original stories. A wide range of genres was represented, including legal thrillers, science-fiction, and romance.

The corpus consists of 1,203,347 tokens and 13,797 types, as counted by Wordsmith Tools 4.0. The number of tokens is the total number of running words in a text while types refers to individual words, counted once. The counts were of unlemmatized words: e.g., *work*, *working* and *works* are taken as different words (types). The texts in the corpus contained between 300 to 2800 headwords, which are suitable for L2 learners from elementary to advanced levels.

Data Cleaning

The graded readers were copied onto Notepad (Version 1903) and saved as .txt files, using UTF-8 encoding. Headers, chapter numbers and page numbers were deleted from the texts. The files were manually checked for errors.

Data Analysis

The corpus was first analysed by the MAT with respect to Biber's (1995) Dimensions. The *Tag and Analyze* function in the programme was used, with the number of tokens set to the number of tokens in the corpus (n = 1,203,347). *All Tags* was selected, resulting in 67 linguistic variables

(41 POS, 24 [clausal] variables, type-token ratio (TTR) and average word length) identified for analysis.

The results of this preliminary study (see Appendix A for a breakdown of how each title in the corpus was classified) evinced some limitations of conducting analyses on a single type of text (fictional texts) using MAT. The majority (81 of 90) of the graded readers were classified as Imaginative Narrative texts. The remaining were classified as General Narrative Exposition (n=5), Informational interaction (n=2) and Involved persuasion (n=2). These exceptions occurred at various graded reader levels (300 to 2500 headwords). MAT classifies text types using the "Euclidean distance from the centroids of the clusters reported in Biber (1989)" (Nini, 2019, pp.10-11). Variables with large sizes are given more weight (Härdle & Simar, 2015) and thus, may disproportionately affect the clustering. Largely, the MAT analysis seemed to characterize the graded readers in the corpus as types of narratives. While accurate, it is evident that analysing linguistic variation within this genre required additional methods.

To address this, the z-scores output from MAT was subjected to LCC analysis. Unlike the Euclidean distance used by MAT, LCC analysis identifies hidden groups in data through the probability that the cases within the dataset belong to a group. Consequently, scaling ceases to be an issue (Vermunt & Magidson, 2002). LCC analysis has the added advantage of model selection criteria and probability classification (Vermunt & Magidson, 2005a). LCC analysis may also be more efficient and more robust although with sufficient large numbers of samples, these differences do not signify much (Chiu, Douglas, & Li, 2009). For the small corpus in this study, however, LCC analysis would be more likely to produce more robust models.

The underlying principle of classification is that subgroups can be observed based on the patterns of occurrence of variables. Distinctive patterns are taken as indicating some latent trait(s) unique to the group, although what these traits are, is subject to interpretation. As applied in this study, texts in the corpus are classified based on the z-scores of the 67 linguistic variables; that is, groups of texts that share similar linguistic features were identified through LCC analysis. Post-hoc corpus analyses were used for interpretation of the clusters.

The z-scores from MAT were input into the Latent Gold software to form latent class cluster models to determine if the linguistic features of the texts form distinct groupings. The z-scores of each linguistic feature are the relative frequencies of the feature standardized to the mean frequencies of the feature in the original corpus of written and spoken texts compiled by Biber (1995) mentioned above (Nini, 2019).

Four cluster models were estimated with the 67 variables as indicators for the models; that is, we explored the usefulness of four types of text classification models (i.e., as a single, two, three and four groups of texts) to determine the optimal way of classifying the texts. To select the best-fitting model among these four models, Goodness-of-Fit statistics were used to evaluate the best fit. We used the commonly used Information Criterion (IC) statistics, which estimate the degree of error in each model (Vermunt & Magidson, 2002). Lower IC statistics indicate a better fit of the model to the data (Burnham & Anderson, 2004). Specifically, we used measures estimated using log-likelihood (LL), namely, the Akaike Information Criterion (AIC), Consistent AIC (CAIC) and Bayesian Information Criterion (BIC). By comparing the fit statistics (AIC, CAIC and BIC) for the models, we would find out whether it is best to classify the graded readers into groups of one, two, three, or four, based on their linguistic features. The selection of the best-fitting model was further verified with the expected misclassification error. This was calculated by the cross classification of the modal classes with the probabilistic classes, with values closest to 0 indicating better fit (Vermunt & Magidson, 2005a). In other words, the optimal model would have a misclassification error equal or close to zero.

After the best-fitting model was selected, the characteristics of each cluster in the model were examined. We examined several features in the best-fitting model. Specifically, we examined the "Model for Indicators" to determine how each linguistic feature behaves under the model. This provided an indication of the extent to which each linguistic feature distinguishes each group of texts. In addition, we examined the Wald statistics and its p-value showing how significant each

variable is. The significance level was set at p = .05, i.e., linguistic features with p <.05 would discriminate the clusters in a statistically significant way. The clusters were further characterized by examining the "Profile Plot" of the significant clausal variables. The Profile Plot is rescaled on a 0 – 1 scale such that the frequencies of the variables are comparable (Vermunt & Magidson, 2005b), thus, indicating the relative importance of each linguistic feature for the cluster of texts.

Next, post-hoc analyses of the clausal variables were carried out to characterize each cluster. The clausal variables encode the linguistic features identified by Biber (1995) by parsing the POS-tags previously assigned by the Stanford Tagger. As such, an analysis of this sub-set provides data-reduction while still allowing for the most pertinent POS to be included for analysis. The clausal variables are indicated in square brackets [ ], following the convention set in MAT. For significant clausal variables (i.e., variables in the best-fitting model with p<.05), the Concordance function in Wordsmith Tools (Version 4.0) was used to compute the occurrences of word + tag or tag + word, in 5L to 5R word clusters. The minimum frequency was set to 5, to ensure only meaningful patterns were identified.

Results

Latent Class Cluster Analysis (LCC)

To determine which linguistic features were meaningful to distinguishing groups within the corpus, a LCC analysis was performed on the z-scores on the 67 variables produced by MAT. Four latent class models were fitted. While the AIC(LL) was the lowest for the four-cluster model, the BIC(LL) and CAIC(LL) values were the lowest for the 3-cluster model (Table 1). As the BIC(LL) statistic tends to be more reliable (Nylund et al., 2007; Vrieze, 2012), the three-cluster model was selected as having the best fit.

Table 1

Goodness-of-Fit Statistics of the LCC Models

Model	BIC(LL)	AIC(LL)	CAIC(LL)	No. of Parameters	Proportion of Classification Errors
1-Cluster	7258.84	6923.87	7392.84	134	0
2-Cluster	5363.77	4691.32	5632.77	269	0
3-Cluster*	**4975.87**	3965.95	**5379.87**	404	0.0005
4-Cluster	5093.90	**3746.50**	5632.90	539	0.0002

Note. The lowest fit index is indicated by bold print. LL: Log-likelihood; BIC: Bayesian information criterion; AIC: Akaike information criterion; CAIC: Consistent Akaike information criterion.

The cross-tabulation of the modal classes with the probabilistic classes (Table 2) confirmed the three-cluster model as best-fitting. In a model with optimal accuracy, all the texts would be classified in only one latent cluster (values on the diagonal). In this three-cluster model, the largest cluster is Cluster 1, with slightly less than half of the texts belonging to it. Clusters 2 and 3 are almost equal in size, with 24 and 22 graded readers respectively. In addition, the estimated miscalculation error values (off-diagonal values, in bold text in Table 2 below) are zero or close to zero, confirming the optimality of the modelling.

Table 2

Cross-Tabulation of Probabilistic and Modal Class Assignment in the 3-Cluster Model

	Modal			
Probabilistic	Cluster 1	Cluster 2	Cluster 3	Total
Cluster 1	43.98	**0**	**0.03**	44.01
Cluster 2	**0**	24	**0**	24
Cluster 3	**0.02**	**0**	21.97	21.99
Total	44	24	22	90

Note. Estimated miscalculation error values are in bold text.

There are far more tokens in the texts in Cluster 3 than that of Cluster 2 despite there being nearly the same number of texts both clusters (Table 3). Further comparisons between clusters are therefore made using normalized figures.

Table 3

Number of Texts and Tokens in Clusters in the 3-Cluster Model

Cluster	Number of Texts	Tokens
1	44	605,169
2	24	131,898
3	22	466,280

Note. Tokens refers to the total number of running words (as computed by Wordsmith Tools).

Although LCC identified group membership, what characterizes each cluster (or group) is not readily observable. The *Models of Indicators* statistics provide insight into which linguistic features differentiate the clusters. In total, 50 out of the 67 variables significantly distinguished the texts in the corpus (p<.05): 31 POS, average word length and 18 clausal variables. The behavior of the clausal variables in the clusters is of particular interest and the results for this subset are presented in Table 4 (see appendix for a complete table). For the 18 significant clausal variables, *Split auxiliaries* [SPAU] had the highest R^2 value (0.69) while *Public verb* [PUBV] had the lowest (0.12); the closer the R^2 value is to 1, the better the predictions (Vermunt & Magidson, 2005a). This means that *Split auxiliaries* [SPAU] are better predictors of whether a text belongs to any of the clusters, as compared to *Public verb* [PUBV]. Further comparisons of the R^2 provided insights into the importance of the linguistic feature in distinguishing groups of graded readers. It should be noted that higher R^2 values indicate that the variable (indicator) is explained better by the model (Aryadoust, 2020). For example, the R^2 index of *Split auxiliaries* [SPAU] is 0.69, which indicates that 69% of the variance in *Split auxiliaries* [SPAU] is explained by the model.

Table 4

Models of Indicators of Clausal Variables in the 3-Cluster Model

Clause [Tag]	Mean of Cluster 1	Mean of Cluster 2	Mean of Cluster 3	Wald	p	R^2
Be as main verb [BEMA]	-0.14	0.37	-0.23	27.08	**1.30E-06**	0.24
By-passives [BYPA]	-0.03	-0.17	0.20	81.66	**1.90E-18**	0.41
Contractions [CONT]	-0.10	0.16	-0.05	3.28	0.19	0.04

Clause [Tag]	Mean of Cluster 1	Mean of Cluster 2	Mean of Cluster 3	Wald	*p*	R^2
Agentless passives [PASS]	0.03	-0.29	0.26	173.56	**2.10E-38**	0.57
Past participial clauses [PASTP]	0.13	-0.13	0.00	2.18	0.34	0.03
Perfect aspect [PEAS]	0.02	-0.97	0.95	375.01	**3.70E-82**	0.62
Pied-piping relative clauses [PIRE]	-0.04	-0.07	0.11	25.38	**3.10E-06**	0.31
Present participial clauses [PRESP]	-0.01	-0.60	0.62	129.96	**6.00E-29**	0.45
Private verbs [PRIV]	0.02	-0.45	0.44	60.28	**8.10E-14**	0.41
Pro-verb do [PROD]	0.03	-0.08	0.06	5.89	0.05	0.07
Public verbs [PUBV]	-0.30	0.58	-0.28	7.91	**0.02**	0.12
Sentence relatives [SERE]	-0.13	-0.43	0.55	30.12	**2.90E-07**	0.31
Seem|appear [SMP]	0.12	-0.82	0.70	278.84	**2.80E-61**	0.57
Split auxiliaries [SPAU]	0.06	-0.69	0.63	230.71	**8.00E-51**	0.69
Split infinitives [SPIN]	-45.48	-45.48	90.96	3.51	0.17	0.11
Stranded preposition [STPR]	0.04	-0.10	0.06	4.26	0.12	0.05
Suasive verbs [SUAV]	-0.02	-0.08	0.10	1.74	0.42	0.02
Subordinator that deletion [THATD]	-0.08	-0.40	0.49	77.62	**1.40E-17**	0.47
WH-clauses [WHCL]	0.23	-0.83	0.61	87.52	**9.90E-20**	0.49
WH-relative clauses on object position [WHOBJ]	-0.01	-0.05	0.06	55.88	**7.30E-13**	0.31
Direct WH-questions [WHQU]	-0.54	2.04	-1.50	35.11	**2.40E-08**	0.34
[WHSUB]	0.13	-0.33	0.20	214.57	**2.60E-47**	0.50
Past participial WHIZ deletion relatives [WZPAST]	0.00	-0.03	0.03	18.72	**8.60E-05**	0.13
Present participial WHIZ deletion relatives [WZPRES]	0.13	-0.45	0.32	130.46	**4.70E-29**	0.37

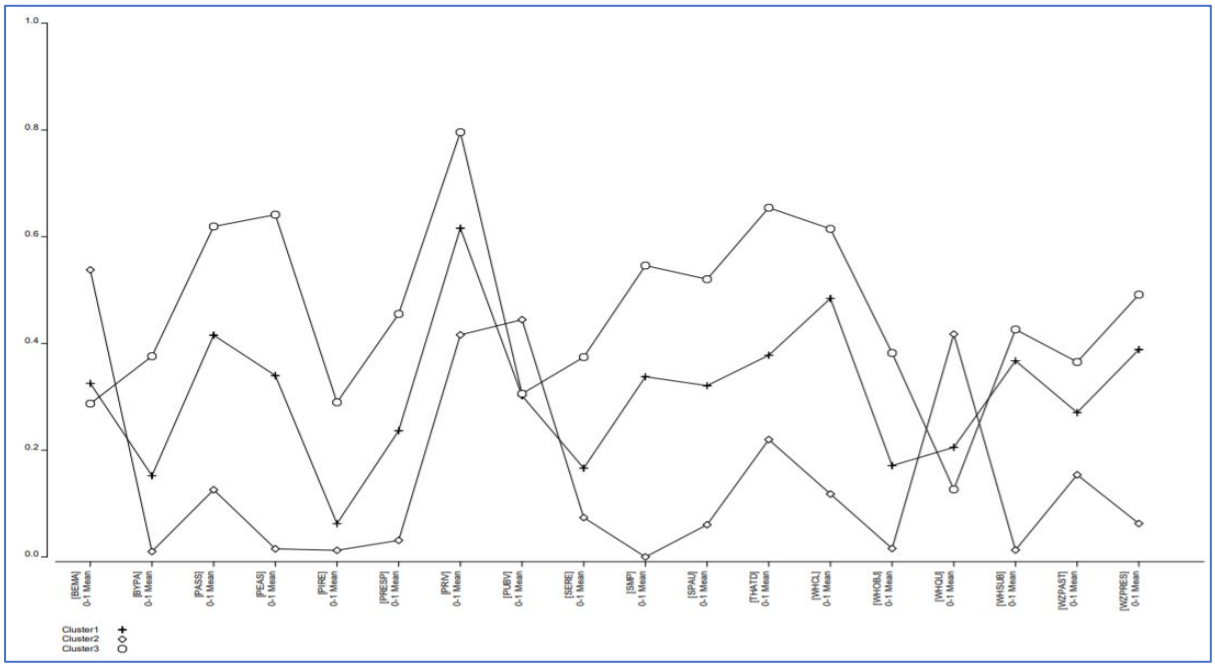
Note. Significant p-values (p<.05) are indicated by bold print.

The patterns of occurrence of these significant clausal variables show the variation in the three clusters (Figure 1). The clusters generally follow the same directional patterns, differing only in magnitude. That is, the relative frequencies of most of these variables were the lowest in Cluster 2, followed by Cluster 1 and finally, Cluster 3. It is important to note that this is irrespective of text length.

The variables that do not follow this pattern were *Be as main verb* [BEMA], *Public verbs* [PUBV] and *Direct WH-questions* [WHQU]. These all had markedly high occurrence in Cluster 2. [BEMA] and [WHQU] occurred more in Cluster 1 than in Cluster 3 while both clusters converged on [PUBV]. The theoretical and pedagogical significance of these patterns is elaborated in the discussion section.

Figure 1

Profile Plot of Clausal Variables in the 3-Cluster Model



Proposed Profile of Clusters in LCC Analysis

The distribution of the texts in the clusters and their headwords provided a preliminary indication of the characteristics of the clusters that emerged in the LCC analysis. Post-hoc concordance analyses of the components of the clausal variables allow a more complete profile of each cluster to be drawn.

The clausal variables mainly vary in magnitude across the clusters, as reflected in Figure 7. The exceptions, *Be as main verb* [BEMA] and *Direct WH-questions* [WHQU], occurred more frequently in Cluster 2. [WHQU] are questions in the combination of

(1) WH-words (*what, when, where*, etc.) + auxiliary verbs.

[BEMA] indicates structures where any form of *be* occurs as a main verb

(2) *Be* {+ negation (XX0)}[1] + determiner (DT)
+ possessive pronoun (PRP$)
+ preposition (PIN)
+ adjective (JJ)

The two structures, *Be as main verb* [BEMA] and *Direct WH-questions* [WHQU], that occur relatively more frequently in Cluster 2 are, arguably, simpler syntactic structures, as compared to, for example, passivation. L2 learners rarely produce passives and relative clauses (McDonough & Trofimovich, 2016). Therefore, the higher relative frequency of these variables in Cluster 2 further supports the supposition that the clustering reflects textual difficulty.

Using Wordsmith Tools, the components of each variable were examined. Figure 2 shows examples of the BEMA structures in Cluster 2. In the centre position is the variable in question (i.e., BEMA). In the L5 to R5 positions, are the words and tags that occur at least 5 times in the clusters. Through an examination of these patterns, recurrent lexico-grammatical patterns were identified.

Figure 2

Patterns of BEMA Concordance in Cluster 2 in the 3-Cluster Model

BEMA_C2.cnc

File Edit View Compute Settings Windows Help

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	NN	IT	NN	WAS	VBD	BEMA	A	PRED	AND	NN	AND
2	DT	HE	S	VPRT	VPRT		VERY	DT	THE	PRED	I
3	I	FPP1	TPP3	IS	CONT		YOU	AMP	CONT	JJ	THE
4	IT	PIT	PIT	ARE	VB		THE	PIN	A	PIN	HE
5	CC	TPP3	M	WERE			IN	XX0	GOOD	DT	SAID
6	TPP3	I	FPP1	BE			NOT	SPP2	I	VBG	OF
7	RB	SHE	WHQU	AM			NT	FPP1	HAPPY	TPP3	MAN
8	HE	TO	RE				MY	RB	IN	RB	SHE
9	JJ	RB	DEMP				AFRAID	TPP3	VERY	FPP1	YOU
10	THAT	DEMP	TO				ALL	JJ	BUT	CC	BUT
11	YOU	SPP2	SPP2				SORRY	QUAN	HE	PHC	TO
12	ANDC	WP	RB				DEAD	CD	WITH	TIME	S
13	WHAT	THEY	WP				ANGRY	EMPH	OF	AMP	FOR
14	FPP1	YOU	CONT				IT	PIT	ANGRY	ANDC	PRED
15	PUBV	THIS	PRMD				HAPPY	DWNT	GOING	SPP2	NOW
16	VBD	FACE	QUPR				AN	DEMO	TO	IN	WITH
17	VBG	WHQU	CC				YOUR	TIME	FOR	VB	IN
18	SHE	WHO	JJ				ON	PLACE	NOW	CD	ABOUT

concordance | collocates | plot | patterns | clusters | filenames | source text | notes

In *Be as main verb* [BEMA], the forms of *be* in Cluster 2 were the present tense forms (*am*, *is*, *are*), the past tense forms (*was*, *were*) and the bare infinitive. Clusters 1 and 3 had, in addition to all the forms in Cluster 2, the participles *been* and *being*. A similar pattern of complexity holds for *Direct WH-questions* [WHQU], where there were more types of modals in Clusters 1 and 3 as compared to Cluster 2 (Table 5).

Table 5

Cluster Distribution of Modals in Direct WH-questions [WHQU] *in the 3-Cluster Model*

Modals	Cluster 1	Cluster 2	Cluster 3
am/are/is *was/were*	x	x	x
does/do *did*	x	x	x
has/have *had*	x		x
Contraction *'s*	x	x	x
Can	x	x	x
Could	x		x
Will	x		x
Would	x		x
shall	x		x
should	x		x
Contraction *'ll*	x		

Note. x indicates that the modal co-occurred at least 5 times with the other components in the structure.

The distribution of variables related to lexis – *Public verbs* [PUBV], *Private verbs* [PRIV], *Seem/appear* [SMP] – lend further credence to this characterization of the clusters. As seen in Table 6 below, Cluster 3 tended to have the greatest number of types for each variable. To account for varying text lengths in each cluster, the normalized TTRs are also given here. TTR is calculated by dividing the number of types and tokens. In this study, TTR is normalized to per 10,000 tokens. For example, for every 10,000 words in the text, *Public verbs* [PUBV] appear 1.16 times in Cluster 3, 0.63 times in Cluster 1 and 0.61 times in Cluster 2.

Table 6

Number of Types and TTR for PUBV, PRIV and SMP in the 3-Cluster Model

		PUBV		PRIV		SMP	
Cluster	Tokens	Types	TTR	Types	TTR	Types	TTR
1	605,169	38	0.63	92	1.52	720	11.90
2	131,898	8	0.61	35	2.65	0	0
3	466,280	54	1.16	106	2.27	720	15.44

Note. TTR = Type Token Ratio (normalized to per 10,000 words)

Moving deeper into syntactic constructions, an examination of *Subordinator that deletion* [THATD] shows the degree of embedding in clauses across the clusters. The [THATD] construction is "a form of syntactic reduction" (Biber, 1995, p. 244), where *that* has a null option. This is not a syntactically free construction (see Dor, 2005, on semantic restrictions; Quirk et al., 1985, on the rules of complementation). The following clausal structures were present in all three clusters (examples from texts in the corpus are given beneath):

(1) Verb [THATD] + demonstrative pronoun
Proximo knew [THATD] *this*_DEMP *meant that the Emperor had* (*Gladiator*)

(2) Verb [THATD] + a subject form of a personal pronoun
She thought [THATD] *he_TPP3 was boring* (*Double Cross*)

Cluster 2 showed the least variation in this variable. Besides (3) and (4), the following occurred in Cluster 2:

(3) Verb [THATD] + determiner (DT) + noun (N) + verb
But don't think_VB [PRIV] [THATD] the_DT Queen_NN understood_VBD [PRIV] him
(The Elephant Man)

For Cluster 1, the structure (5) is extended in the types of determiners after the verb. In this cluster, the quantifiers (QUAN) *many, several, some* and *any,* and cardinal numbers (CD) were present alongside the determiners in Cluster 2. An additional structure was present in Clusters 1 and 3:

(4) Verb + adjective (JJ) + noun (N) + verb (V)
I see_VPRT [PRIV] [THATD] healthy_JJ children_NN running_VBG [WZPRES] *around* (*The Mosquito Coast*)

Thus far, Cluster 2 demonstrably had simpler clausal structures and more restricted lexis within the clausal variables. The differences between Cluster 1 and Cluster 3 were less obvious; there was some variation of lexical items, where Cluster 3 tended to have more varied lexis. In terms of the clausal structures, Cluster 3 tended to include a wider variety of POS.

In *WH-relative clauses on object position* [WHOBJ], nominalization (NOMZ) occurred in Cluster 3 and not in Cluster 1. Differences in lexis between clusters were repeated in [WHOBJ], where there was only a single occurrence of this structure in Cluster 2 (Table 7).

Table 7

WH-Relative Clauses on Object Position [WHOBJ] *in the 3-Cluster Model*

Cluster	any word that is NOT a form of the words ASK or TELL	+ any WH word	+ noun (N),	+ any word NOT RB, XX0, MD, or forms of HAVE, BE or DO
Cluster 1	NN	which who	I / you / he / she	Had
Cluster 2	There is only one occurrence of this structure.			
Cluster 3	NN **NOMZ**	which who **whom** **whose**	I / you / he / she	**Could** had **was**

Note. Features unique to a cluster are highlighted by **bold** print. MD =modal, NN = nouns, NOMZ = nominalizations, RB = adverb, XX0 = negation.

Turning next to *Split auxiliaries* [SPAU], and its associated variables, the clusters displayed more marked differences in complexity of clausal structures. [SPAU] tend to co-occur with *By-passives* [BYPA] and *Agentless passives* [PASS] (Biber, 1986, p. 393), and, thus, will be discussed together.

There was a wider variety of auxiliary verbs and adverbs used in the [SPAU] construction in Cluster 3, as reflected by the highest numbers of types and TTRs (Table 8) of the component words in the construction.

Table 8

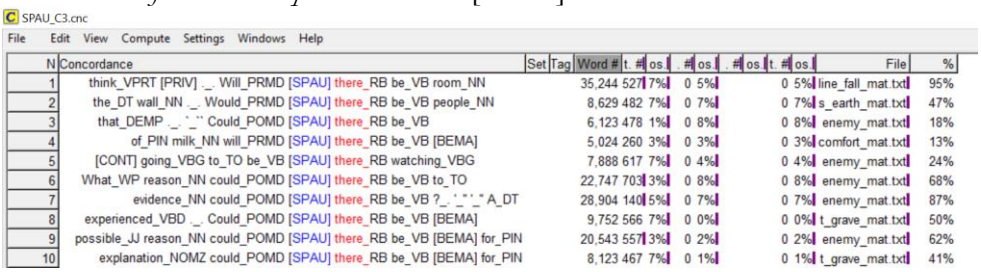
Component Words of [SPAU] *in the 3-Cluster Model*

	Auxiliary		+ Adverb	
Cluster	Type	TTR	Type	TTR
1	15	0.25	26	0.43
2	4	0.30	4	0.30
3	19	0.41	42	0.90

An additional pattern that occurred only in Cluster 3 is the use of the adverb *there*. An examination of the concordance lines (Figure 3) showed that *Split auxiliaries* [SPAU] with the word *there* occurred with the word *be* to denote probability.

Figure 3

Concordance of There *in Split Auxiliaries* [SPAU]



SPAU_C3.cnc

File Edit View Compute Settings Windows Help

N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	think_VPRT [PRIV] _. Will_PRMD [SPAU] there_RB be_VB room_NN			35,244	527	7%	0	5%			0	5%	line_fall_mat.txt	95%
2	the_DT wall_NN _. Would_PRMD [SPAU] there_RB be_VB people_NN			8,629	482	7%	0	7%			0	7%	s_earth_mat.txt	47%
3	that_DEMP _. ``_`` Could_POMD [SPAU] there_RB be_VB			6,123	478	1%	0	8%			0	8%	enemy_mat.txt	18%
4	of_PIN milk_NN will_PRMD [SPAU] there_RB be_VB [BEMA]			5,024	260	3%	0	3%			0	3%	comfort_mat.txt	13%
5	[CONT] going_VBG to_TO be_VB [SPAU] there_RB watching_VBG			7,888	617	7%	0	4%			0	4%	enemy_mat.txt	24%
6	What_WP reason_NN could_POMD [SPAU] there_RB be_VB to_TO			22,747	703	3%	0	8%			0	8%	enemy_mat.txt	68%
7	evidence_NN could_POMD [SPAU] there_RB be_VB ?_. ``_`` ``_`` A_DT			28,904	140	5%	0	7%			0	7%	enemy_mat.txt	87%
8	experienced_VBD _. Could_POMD [SPAU] there_RB be_VB [BEMA]			9,752	566	7%	0	0%			0	0%	t_grave_mat.txt	50%
9	possible_JJ reason_NN could_POMD [SPAU] there_RB be_VB [BEMA] for_PIN			20,543	557	3%	0	2%			0	2%	enemy_mat.txt	62%
10	explanation_NOMZ could_POMD [SPAU] there_RB be_VB [BEMA] for_PIN			8,123	467	7%	0	1%			0	1%	t_grave_mat.txt	41%

The *Agentless passives* [PASS][2] and *By-passives* [BYPA] constructions not only co-occurred with *Split Auxiliaries* [SPAU]; [SPAU] is a sub-set of both. Adding the preposition *by* to the end of [PASS] results in the by-passives [BYPA], and thus, can be considered as part of the [PASS] structure. A similar distribution in lexis was found across the clusters (Table 9) but of more pertinence here is the structure of the [PASS] clauses.

Table 9

Agentless Passives [PASS] in the 3-Cluster Model

Cluster	Any form of *Be*		{+ adverb (RB)}		+ past (VBD) or participle (VBN) form of verb	
	Types	TTR	Types	TTR	Types	TTR
1	9	0.15	13	0.21	87	1.44
2	There are too few occurrences to form a pattern.					
3	11	0.24	16	0.34	105	2.25

The only passive structure in Cluster 2 is:

(5) Any form of *Be* {+ adverb (RB)} + past or participle form of verb (VBN or VBD)

When the optional adverb in (5) occurred, a split auxiliary is used. An example from *Death in the Freezer* is given to illustrate:

*Dan Future's hair was*_VBD [SPAU] [PASS] *beautifully*_RB *cut*_VBN

In Clusters 1 and 3, (5) is extended to include a nominal form (a noun or personal pronoun) between *be* and the participle:

(6) Any form of *Be* + **a nominal form** + past or participle form of verb (VBN or VBD)

*That there were*_VBD[PASS] *names*_NN *written*_VBN[PUBV] (*Wuthering Heights*)

A similar pattern of variation was found in the componential lexis of *Pied-piping clauses* [PIRE], *Sentence relative* [SERE] and *Present participial clauses* [PRESP]. Variation in complexity of tense-aspect was observed across the clusters in *Perfect aspect* [PEAS]. The levels of complexity in tense-aspect increases from Cluster 2 to Cluster 1 and subsequently, Cluster 3. Only the present perfect was present in Cluster 2:

(7) *have* + contraction *n't* + *got*

*You have*_VPRT [PEAS] *n't*_XX0 [CONT] *got*_VBN *any money*.

(*The Lottery Winner*)

Alongside the present perfect, the past perfect was in Clusters 1 and 3:

(8) *has* / *have* / *had* + adverb (RB) or negation + past or participle form of verb (VBD/VBN)

*knowing him had*_VBD [SPAU] [PEAS] *also*_RB *damaged*_VBN *her marriage*

(*Dr Zhivago*)

Cluster 3, in addition, had the present perfect continuous:

(9) *having* + {adverb (RB) or negation} + past or participle form of verb (VBD/VBN)

*Having*_VBG [PEAS] *amused*_VBN *himself*

(*Pride and Prejudice*)

There were fewer occurrences of interrogatives in the perfect aspect, as compared to affirmatives, in all clusters. Consequently, there were insufficient occurrences adverbs or negation for patterning. Some variables ([WHSUB], [WZPRES], and [WZPAST]) did not conform exactly to this pattern of syntactic or lexical complexity, for Clusters 1 and 3.

Discussion

The results of the preliminary analysis did not reveal much in terms of variation within this corpus. Additionally, the classification by MAT of the corpus into four text types – *Imaginative narrative* (n=81), *General narrative exposition* (n=5), *Informational interaction* (n=2) and *Involved persuasion* (n=2) – is very uneven in terms of group sizes. In comparison, the LCC analysis produced a more meaningful clustering. From a methodological perspective, the utility of LCC analysis for modelling smaller-scale data is underscored in this study, pointing to the feasibility of using the method to augment the customary methods used in corpus linguistics. Researchers looking to studying small corpora and single genres might consider using LCC analysis as a way of discovering latent factors in their data.

An examination of the headwords and the clusters revealed a striking correspondence between headwords and Cluster 2; graded readers on the lower end of headwords count tended to be placed in Cluster 2. The headwords in Cluster 2 ranged from 300 to 700, with some variation among the books with 700 headwords (OUP Level 2; see Appendix). Even taking into account the different ways publishers may have used to count headwords, this distinct pattern probably points to variation in the complexity of the linguistic features that form the basis of clustering. The division between Cluster 1 and Cluster 3 is less distinct. Cluster 1 seems to be of mid-level difficulty (700 to 2500 headwords) and texts in Cluster 3 mostly fall on the upper ends of the headwords levels (1400 to 2800 headwords).

Interpretation of Clusters of Graded Readers

What has emerged from the study of the three clusters is that the language in graded readers does not only increase in lexical variation as the levels increase, but also shows progression in syntactic complexity. Cluster 2 seems to be the 'basal level', having the least variation in lexis within the clausal structures identified and a higher relative frequency of simpler structures like *WH-questions* and *Be as a main verb*. Increasing embedding of clauses was seen in Clusters 1 and 3, as compared to Cluster 2. Clusters 1 and 3 were mostly distinguished by differences in lexis within the clausal structures, with Cluster 3 having a wider range of lexis associated with the clausal variables. Cluster 1 may be appropriately designated 'transitional', given the degree of overlap between it and Cluster 3. Other studies have found that texts at this level are difficult to classify, attributable to the "transitory nature of the level" (Crossley et al., 2011, p. 97). In keeping with most level structures, Cluster 3 may be referred to as 'advanced'. The pedagogical implications of the results of the LCC analysis, therefore, are perhaps more profound for novice readers, rather than for more advanced readers. For ER, the provision of a wide variety of reading materials that students can comprehend is of great importance. Teachers of beginner level students, in particular, should take the care to provide fictional texts that are less abstract with more dialogue and other easy-to-understand rhetorical features, like direct WH-questions, of which there is greater abundance in beginner texts in the 300 to 700 headwords range. Texts with an abundance of more complex structures may be reserved for when students progress to a higher level. Teachers may also refer Table 4 as a general guide to the types of linguistic features that are more likely to distinguish text complexity (features with higher R^2 values, e.g., split auxiliaries, perfect aspect, agentless passives).

The results of this study have also shown that restricting lexis for text simplification go hand-in-hand with a degree of syntactic simplification. For example, although the word family of *have* is present at the basal level (Cluster 2), the absence of other tense-aspects besides the present

perfect in the cluster indicates the degree of syntactic simplification that concurrently occurs, whether by design or chance, with lexical simplification. This seems to suggest that teaching grammar and vocabulary should go hand-in-hand. In his Four Strands, Nation (2007) calls for a balance of form-focused instruction, meaning-focused input, meaning-focused output and fluency development. For meaning-focused input and fluency development, the selection of appropriate texts by teachers, as suggested above, may be enhanced by knowledge of the characteristics of the various levels of texts. Additionally, since ER provides meaning-focused input and fluency development, teachers may augment these two strands with form-focused instruction that focuses on the forms already present in the graded readers that their students are reading. In this way, in-class learning activities may reinforce the grammar and vocabulary learnt implicitly through ER. The lists of grammatical structures and the associated vocabulary produced by this study, therefore, may be useful for teachers planning such form-focused instruction.

The interpretation of these clusters, derived as they were from frequency counts and statistical modelling, does take a position of usage-based acquisition. The lack of consensus amongst researchers on the processes of language acquisition or learning precludes absolute certitude in this position but, from a pedagogical perspective, it is hard to argue against the benefits of repetition and reinforcement. In this, the lists of grammatical structures and associated lexis in graded readers presented in this study[3] may provide a guide for teachers wishing to supplement instruction with ER (or vice-versa).

Limitations

It must be acknowledged that a more balanced corpus that incorporates other text types would provide a more comprehensive profile of language learner literature. The use of a restricted genre as a study corpus also highlighted the lack of variation in terms of Dimensions. This is perhaps an indication that learners may not be exposed to a full range of linguistic variation should they read only fictional graded readers. The inclusion of other types of texts in language learner literature would be valuable, especially if such a study were to go beyond an examination of vocabulary alone.

Conclusion

This study set out to provide a more complete linguistic profile of graded readers that accounts for both lexis and grammar. We found that variation in the language of graded readers largely coincided with the headwords levels listed by publishers, suggesting that headwords levels seem to also indicate variation in grammatical complexity. An argument can be made that progression in complexity – both lexical and syntactic – of linguistic input is present at each graded reader level.

The lists of structures and lexis accompanying this study may also be of pedagogical use to teachers. As the clusters display progressions in complexity, the usefulness of the lists is enhanced; teachers may accordingly draw from the lists the forms that are most suitable for their class. Thus, the results of this study may provide teachers the resources to plan for data-driven learning without having to build their own corpora.

Although, in the present study, a precise demarcation cannot be established for more advanced students, further research may determine the boundary separating transitionary texts from advanced texts. It may very well prove to be that such a boundary is non-existent; at higher levels, teachers may guide students to self-select texts more independently instead.

Grammatical complexity seemed to be related to lexical variability, suggesting that teaching grammar and vocabulary should go together. As this study has demonstrated, taking grammar into account is essential to gaining a more accurate understanding of language. While this study has demonstrated that lexis has a role in grammatical complexity, the extent to which this relationship

is significant was not examined. Forthcoming research in this direction will enrich both theory and practice.

Acknowledgements

This study is based on the first author's MA dissertation, completed at National Institute of Education, Nanyang Technological University, under the supervision of the second and third authors. This study has not been funded by any institution and, therefore, the authors declare no conflict of interest.

The authors thank Professor Tom Cobb, for permission to use the corpus of graded readers from the Lextutor website (https://www.lextutor.ca/).

About the Authors

Azrifah Zakaria: A PhD candidate at the National Institute of Education, Nanyang Technological University, Singapore. Her research interests are in corpus linguistics, language assessment and computer assisted language learning. Besides teaching, she has previously worked in early childhood education and intervention.

Willy A Renandya: A language teacher educator with extensive teaching experience in Asia. He currently teaches applied linguistics courses at the National Institute of Education, Nanyang Technological University. He has given numerous keynote presentations at international ELT conferences and has published extensively in the area of second language education.

Vahid Aryadoust: An Associate Professor of language assessment at the National Institute of Education of Nanyang Technological University. Vahid has published his research in Language Testing, Language Assessment Quarterly, Assessing Writing, Educational Assessment, Educational Psychology, and Computer Assisted Language Learning, etc. He has also (co)authored multiple book chapters and books published by Routledge, Cambridge University Press, Springer, Cambridge Scholar Publishing, Wiley Blackwell, etc. He teaches graduate courses on Oracy, language assessment, and research methods.

Endnotes

[1] Optional elements are indicated by curly brackets {}.

[2] Occurrences of the word pass in Cluster 1 (n=26) and Cluster 3 (n=33) were excluded from analysis. There were no occurrences in Cluster 2.

[3] The full analyses are available from authors upon request.

References

Allan, R. (2016). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System*, *59*, 61–72. https://doi.org/10.1016/j.system.2016.04.005

Aryadoust, V. (2020). Measureable dimensions of visual mental imagery and their relationship with listening comprehension: Evidence from forensic arts and latent class analysis. *Imagination, Cognition and Personality, 39*(3), 291–319. https://doi.org/10.1177/0276236619829879

Bamford, J., & Day, R. R. (2004). *Extensive reading activities for teaching language*. Cambridge University Press.

Berber Sardinha, T., Veirano Pinto, M., & Berserik, F. (Eds.). (2015). *Multi-dimensional analysis, 25 years on a tribute to Douglas Biber*. [electronic resource]. John Benjamins Publishing Company.

Berber Sardinha, T., & Veirano Pinto, M. (2019). Dimensions of variation across American television registers. *International Journal of Corpus Linguistics*, *24*(1), 3–32. https://doi.org/10.1075/ijcl.15014.ber

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*(2), 384–414. https://doi.org/10.2307/414678

Biber, D. (1995). *Variation across speech and writing*. Cambridge University Press. (Original work published 1988).

Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1), 3–43. https://doi.org/10.1515/ling.1989.27.1.3

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics*, *15*(2), 223–250. https://doi.org/10.1017/S1360674311000025

Burnham, K., & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633–665. https://doi.org/10.1007/s11336-009-9125-0

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101. http://nflrc.hawaii.edu/rfl/April2011/articles/crossley.pdf

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*(1), 89-108. http://dx.doi.org/10.1177/1362168811423456

Crossley, S. A., Yang, H.S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language, 26*(1), 92–113. http://nflrc.hawaii.edu/rfl/April2014/articles/crossley.pdf

Dor, D. (2005). Toward a semantic account of that-deletion in English. *Linguistics, 43*(2), 345–. https://doi.org/10.1515/ling.2005.43.2.345

Ellis, N.C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Friginal, E., Pearson, P., Di Ferrante, L., Pickering, L., & Bruce, C. (2013). Linguistic characteristics of AAC discourse in the workplace. *Discourse Studies*, *15*(3), 279–298. https://doi.org/10.1177/1461445613480586

Graesser, A., McNamara, D., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210–229. https://doi.org/10.1086/678293

Gray, B. (2013). More than discipline: uncovering multi-dimensional patterns of variation in academic research articles. *Corpora*, *8*(2), 153–181. https://doi.org/10.3366/cor.2013.0039

Härdle, W., & Simar, L. (2015). *Applied multivariate statistical analysis*. (4th ed.). Springer.

Jeon, E.Y., & Day, R. R. (2016). The effectiveness of ER on reading proficiency: A meta-analysis. *Reading in a Foreign Language, 28*(2), 246–265. http://www.nflrc.hawaii.edu/rfl/October2016/articles/jeon.pdf

Kano, M. (2015). Revealing factors affecting learners' sense of "difficulty" in extensive reading through reader corpora. *Procedia – Social and Behavioral Sciences, 198*, 211-217. https://doi.org/10.1016/j.sbspro.2015.07.438

Latent Gold (Version 4.5) [Computer software]. Available at: https://www.statisticalinnovations.com/

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. https://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33(3), 299–320. https://doi.org/10.1093/applin/ams010

McDonough, K., & Trofimovich, P. (2016). Structural priming and the acquisition of novel form-meaning mappings. In T. Cadierno, S. Eskildsen, & A. Barraja-Rohan (Eds.). *Usage-based perspectives on second language learning [electronic resource]*. De Gruyter Mouton.

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly, 49*(1), 6–37. https://doi.org/10.1002/tesq.157

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching, 1*(1), 2–13. https://doi.org/10.2167/illt039.0

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63,* 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. & Waring, R. (2020). *Teaching extensive reading in another language*. Routledge.

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In T. Berber Sardinha & M. Veirano Pinto (Eds.). *Multi-dimensional analysis: Research methods and current issues,* 67-94. Bloomsbury Academic. Advance online publication. https://niniandrea.files.wordpress.com/2019/06/pre-print-the-multidimensional-analysis-tagger.pdf

Nini, A. (2020). Multidimensional Analysis Tagger (Version 1.3.1) [Computer software]. Available at: https://sites.google.com/site/multidimensionaltagger/versions

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modelling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Vermunt, J.K. & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A. McCutcheon (Eds.). *Applied latent class analysis [electronic resource]*. Cambridge University Press.

Vermunt, J.K. & Magidson, J. (2005a). *Latent Gold 4.0 user's guide*. Statistical Innovations Inc. Retrieved from: https://www.statisticalinnovations.com/wp-content/uploads/LGusersguide.pdf

Vermunt, J.K. & Magidson, J. (2005b). *Technical guide for Latent GOLD Choice 4.0: Basic and advanced*. Statistical Innovations Inc. Retrieved from: https://www.statisticalinnovations.com/wp-content/uploads/LGCtechnical.pdf

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127

Wan-a-rom, U. (2008). Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, *20*(1), 43–69. https://nflrc.hawaii.edu/rfl/April2008/wanarom/wanarom.pdf

Wordsmith Tools (Version 4.0) [Computer software]. Lexical Analysis Software and Oxford University Press. Available at: https://www.lexically.net/wordsmith/version4/

Appendix A

Classification

The classification of the books according to the different methods (MAT text type classification, and LCC clustering into three and four groups of texts, respectively, in the 3-cluster and 4-cluster solutions) are all presented to highlight the differences (or similarities) between the methods of classification.

Table A1

Classification of Penguin Readers (PP)

Title	Publisher Word Level	Headwords	Text type Classification (MAT)	Cluster # for 3 cluster solution	Cluster # for 4 cluster solution
Brazil 500 years: Voyage to Terra Papagalis	Level 1	300	*General Narrative Exposition*	2	3
Island for Sale	Level 1	300	Imaginative Narrative	2	3
Little Women	Level 1	300	Imaginative Narrative	2	3
Rip Van Winkle and The Legend of Sleepy Hollow	Level 1	300	Imaginative Narrative	2	3
King Arthur and the Knights of the Round Table	Level 2	600	Imaginative Narrative	2	3
Three Short Stories of Sherlock Holmes	Level 2	600	Imaginative Narrative	2	3
Anne of Green Gables	Level 2	600	Imaginative Narrative	2	3
Gucci: Business in Fashion	Level 2	600	*General Narrative Exposition*	2	3
Chance of a Lifetime	Level 3	1200	*Involved persuasion*	1	1
Dangerous Game	Level 3	1200	Imaginative Narrative	1	1
The Fall of the House of Usher and other stories	Level 3	1200	Imaginative Narrative	1	1
Titanic!	Level 3	1200	*General Narrative Exposition*	1	1
Frankenstein	Level 3	1200	*Involved persuasion*	1	1

The Horse Whisperer	Level 3	1200	Imaginative Narrative	1	2
The Count of Monte Cristo	Level 3	1200	Imaginative Narrative	1	1
Sense & Sensibility	Level 3	1200	Imaginative Narrative	1	1
The Ring	Level 3	1200	Imaginative Narrative	1	1
The Picture of Dorian Gray	Level 4	1700	Imaginative Narrative	3	4
The Gladiator	Level 4	1700	Imaginative Narrative	1	1
The Mosquito Coast	Level 4	1700	Imaginative Narrative	1	4
Oliver Twist	Level 4	1700	Imaginative Narrative	1	1
The Client	Level 4	1700	Imaginative Narrative	1	1
The Adventures of Sherlock Holmes	Level 5	2100	Imaginative Narrative	3	4
Dr. Zhivago	Level 5	2100	*General Narrative Exposition*	3	2

Note. Texts *not* classified as Imaginative Narrative are italicized.

Table A2

Classification of Oxford Bookworms (OUP)

Title	Publisher Word Level	Headwords	Text type Classification (MAT)	Cluster # for 3 cluster solution	Cluster # for 4 cluster solution
The Elephant Man	Level 1	400	Imaginative Narrative	2	3
Goodbye Mr Hollywood	Level 1	400	Imaginative Narrative	2	3
The Lottery Winner	Level 1	400	Imaginative Narrative	2	3
Remember Miranda	Level 1	400	Imaginative Narrative	2	3
Mutiny on the bounty	Level 1	400	Imaginative Narrative	2	3
One way ticket	Level 1	400	Imaginative Narrative	2	3
Christmas in Prague	Level 1	400	Imaginative Narrative	2	3

The President's Murderer	Level 1	400	Imaginative Narrative	2	3
The Little Princess	Level 1	400	Imaginative Narrative	2	3
White Death	Level 1	400	*Informational Interaction*	2	3
The Witches of Pendle	Level 1	400	Imaginative Narrative	2	3
Agatha Christie, Woman of mystery	Level 2	700	Imaginative Narrative	1	1
Dead Man's Island	Level 2	700	Imaginative Narrative	2	1
Dracula	Level 2	700	Imaginative Narrative	1	1
Death in the Freezer	Level 2	700	*Informational Interaction*	2	1
Grace Darling	Level 2	700	Imaginative Narrative	2	1
Henry VIII and his six wives	Level 2	700	Imaginative Narrative	1	1
The Jungle Book	Level 2	700	Imaginative Narrative	1	1
Matty Doolin	Level 2	700	Imaginative Narrative	1	1
New Yorkers	Level 2	700	Imaginative Narrative	1	1
Sherlock Holmes: Short Stories	Level 2	700	Imaginative Narrative	1	1
The Death of Karen Silkwood	Level 2	700	Imaginative Narrative	1	1
Chemical secret	Level 3	1000	Imaginative Narrative	1	1
The Picture of Dorian Gray	Level 3	1000	Imaginative Narrative	1	1
Ethan Frome	Level 3	1000	Imaginative Narrative	1	2
Love Story	Level 3	1000	Imaginative Narrative	1	1
Tales of Mystery and Imagination	Level 3	1000	Imaginative Narrative	1	4
Who Sir? Me Sir?	Level 3	1000	Imaginative Narrative	1	1
Wyatt's Hurricane	Level 3	1000	Imaginative Narrative	1	1
The 39 steps	Level 4	1400	Imaginative Narrative	1	4
Cranford	Level 4	1400	Imaginative Narrative	3	2
Lord Jim	Level 4	1400	Imaginative Narrative	3	2
The Moonspinners	Level 4	1400	Imaginative Narrative	1	4
Reflex	Level 4	1400	Imaginative Narrative	1	4
Silas Marner	Level 4	1400	Imaginative Narrative	3	2
The Songs of Distant Earth	Level 4	1400	Imaginative Narrative	3	2
Three Men in a Boat	Level 4	1400	Imaginative Narrative	1	4
The Unquiet Grave	Level 4	1400	Imaginative Narrative	3	2

Washington Square	Level 4	1400	Imaginative Narrative	3	4
Do Androids Dream of Electric Sheep?	Level 5	1800	Imaginative Narrative	1	2
Brat Farrar	Level 5	1800	Imaginative Narrative	3	2
The Bride Price	Level 5	1800	Imaginative Narrative	1	2
Deadlock	Level 5	1800	Imaginative Narrative	1	2
The Garden Party and Other Stories	Level 5	1800	Imaginative Narrative	1	2
Ghost Stories	Level 5	1800	Imaginative Narrative	1	2
The Dead Of Jericho	Level 5	1800	Imaginative Narrative	3	2
Wuthering Heights	Level 5	1800	Imaginative Narrative	3	2
American Crime Stories	Level 6	2500	Imaginative Narrative	3	2
Cold Comfort Farm	Level 6	2500	Imaginative Narrative	3	2
Cry Freedom	Level 6	2500	*General Narrative Exposition*	1	2
Decline and Fall	Level 6	2500	Imaginative Narrative	3	2
The Enemy	Level 6	2500	Imaginative Narrative	3	2
Jane Eyre	Level 6	2500	Imaginative Narrative	3	2
Meteor	Level 6	2500	Imaginative Narrative	3	2
Pride and Prejudice	Level 6	2500	Imaginative Narrative	3	2

Note. Texts *not* classified as Imaginative Narrative are italicized.

Table A3

Classification of Cambridge English Readers (CUP)

Title	Publisher Word Level	Headwords	Text type Classification (MAT)	Cluster # for 3 cluster solution	Cluster # for 4 cluster solution
Inspector Logan	Level 1	400	Imaginative Narrative	2	1
Parallel	Level 1	400	Imaginative Narrative	2	3
How I Met Myself	Level 3	1300	Imaginative Narrative	1	4
Double Cross	Level 3	1300	Imaginative Narrative	1	1
The Ironing Man	Level 3	1300	Imaginative Narrative	1	2
Two Lives	Level 3	1300	Imaginative Narrative	3	4
Nothing but the Truth	Level 4	-	Imaginative Narrative	1	1
Staying Together	Level 4	-	Imaginative Narrative	1	2
A Matter of Chance	Level 4	1900	Imaginative Narrative	1	2

When Summer Comes	Level 4	1900	Imaginative Narrative	3	2
Jungle Love	Level 5	2800	Imaginative Narrative	3	4

Appendix B

Models for Indicators

Tag	Cluster 1	Cluster 2	Cluster 3	Wald	p-value	R^2
AMP	-0.28	0.57	-0.29	7.53	0.023	0.12
ANDC	-0.20	0.32	-0.12	3.13	0.21	0.04
AWL	0.02	-0.21	0.19	18.75	8.50E-05	0.18
CAUS	-0.14	0.36	-0.22	10.93	0.0042	0.17
CONC	-0.07	-0.35	0.42	97.96	5.40E-22	0.45
COND	0.10	-0.68	0.57	284.83	1.40E-62	0.63
CONJ	-0.07	-0.21	0.28	74.50	6.60E-17	0.45
DEMO	-0.03	-0.10	0.13	8.23	0.016	0.06
DEMP	-0.06	-0.05	0.11	5.48	0.065	0.06
DPAR	-0.06	-0.03	0.08	2.77	0.25	0.03
DWNT	0.05	-0.46	0.41	49.19	2.10E-11	0.34
EMPH	0.07	-0.42	0.34	28.85	5.40E-07	0.25
EX	-0.01	0.34	-0.34	9.76	0.0076	0.11
FPP1	0.08	-0.02	-0.06	0.43	0.81	0.00
GER	0.11	-0.32	0.21	32.81	7.50E-08	0.21
HDG	-0.07	0.17	-0.10	2.56	0.28	0.05
INPR	0.01	0.07	-0.08	2.54	0.28	0.02
JJ	0.09	-0.24	0.15	11.42	0.0033	0.13
NEMD	0.00	-0.38	0.38	26.77	1.50E-06	0.25
NN	-0.19	0.84	-0.64	47.47	4.90E-11	0.37
NOMZ	-0.01	-0.14	0.15	41.33	1.10E-09	0.32
OSUB	0.04	-0.68	0.64	265.01	2.80E-58	0.66
PHC	0.06	0.32	-0.38	10.09	0.0065	0.08
PIN	0.02	-0.07	0.05	1.41	0.49	0.02
PIT	0.11	-0.25	0.14	17.63	0.00015	0.11
PLACE	0.10	0.10	-0.20	5.36	0.069	0.04
POMD	-0.06	0.12	-0.06	1.47	0.48	0.03
PRED	-0.22	0.53	-0.31	6.80	0.033	0.08
PRMD	0.18	-0.98	0.81	83.41	7.70E-19	0.53
RB	0.09	-0.31	0.22	29.15	4.70E-07	0.30
SPP2	-0.1678	0.0935	0.0743	7.3074	0.026	0.0751
SYNE	-0.12	-0.03	0.15	8.17	0.017	0.03
THAC	0.12	-0.57	0.46	138.55	8.20E-31	0.42
THVC	0.20	-0.57	0.38	138.20	9.80E-31	0.44
TIME	0.00	0.21	-0.21	10.85	0.0044	0.08
TO	0.07	-0.71	0.64	73.00	1.40E-16	0.48
TOBJ	0.14	-0.44	0.30	164.40	2.00E-36	0.42
TPP3	-0.04	0.00	0.05	0.30	0.86	0.00

TSUB	0.02	-0.34	0.32	49.27	2.00E-11	0.24
TTR	0.00	0.00	0.00	0.00	1	0.00
VBD	0.20	0.09	-0.29	31.82	1.20E-07	0.13
VPRT	-0.19	0.27	-0.08	9.15	0.01	0.14
XX0	-0.23	0.25	-0.02	10.98	0.0041	0.13
[BEMA]	-0.14	0.37	-0.23	27.08	1.30E-06	0.24
[BYPA]	-0.03	-0.17	0.20	81.66	1.90E-18	0.41
[CONT]	-0.10	0.16	-0.05	3.28	0.19	0.04
[PASS]	0.03	-0.29	0.26	173.56	2.10E-38	0.57
[PASTP]	0.13	-0.13	0.00	2.18	0.34	0.03
[PEAS]	0.02	-0.97	0.95	375.01	3.70E-82	0.62
[PIRE]	-0.04	-0.07	0.11	25.38	3.10E-06	0.31
[PRESP]	-0.01	-0.60	0.62	129.96	6.00E-29	0.45
[PRIV]	0.02	-0.45	0.44	60.28	8.10E-14	0.41
[PROD]	0.03	-0.08	0.06	5.89	0.053	0.07
[PUBV]	-0.30	0.58	-0.28	7.91	0.019	0.12
[SERE]	-0.13	-0.43	0.55	30.12	2.90E-07	0.31
[SMP]	0.12	-0.82	0.70	278.84	2.80E-61	0.57
[SPAU]	0.06	-0.69	0.63	230.71	8.00E-51	0.69
[SPIN]	-45.48	-45.48	90.96	3.51	0.17	0.11
[STPR]	0.04	-0.10	0.06	4.26	0.12	0.05
[SUAV]	-0.02	-0.08	0.10	1.74	0.42	0.02
[THATD]	-0.08	-0.40	0.49	77.62	1.40E-17	0.47
[WHCL]	0.23	-0.83	0.61	87.52	9.90E-20	0.49
[WHOBJ]	-0.01	-0.05	0.06	55.88	7.30E-13	0.31
[WHQU]	-0.54	2.04	-1.50	35.11	2.40E-08	0.34
[WHSUB]	0.13	-0.33	0.20	214.57	2.60E-47	0.50
[WZPAST]	0.00	-0.03	0.03	18.72	8.60E-05	0.13
[WZPRES]	0.13	-0.45	0.32	130.46	4.70E-29	0.37