

Supplementary Materials

This is supplementary material for the paper “The impact of lexical bundle length on L2 oral proficiency”.

Table S1.

Assumption Checks for Regression Models

Outcome Variable	Shapiro-Wilk p-value	Linearity (Visual inspection)	Homoscedasticity (Visual inspection)	Multicollinearity (Max VIF)
Presentation Raw Scores	0.04	Not Violated	Not Violated	7.38

Note. See Table S2 below for all VIF values.

Table S2

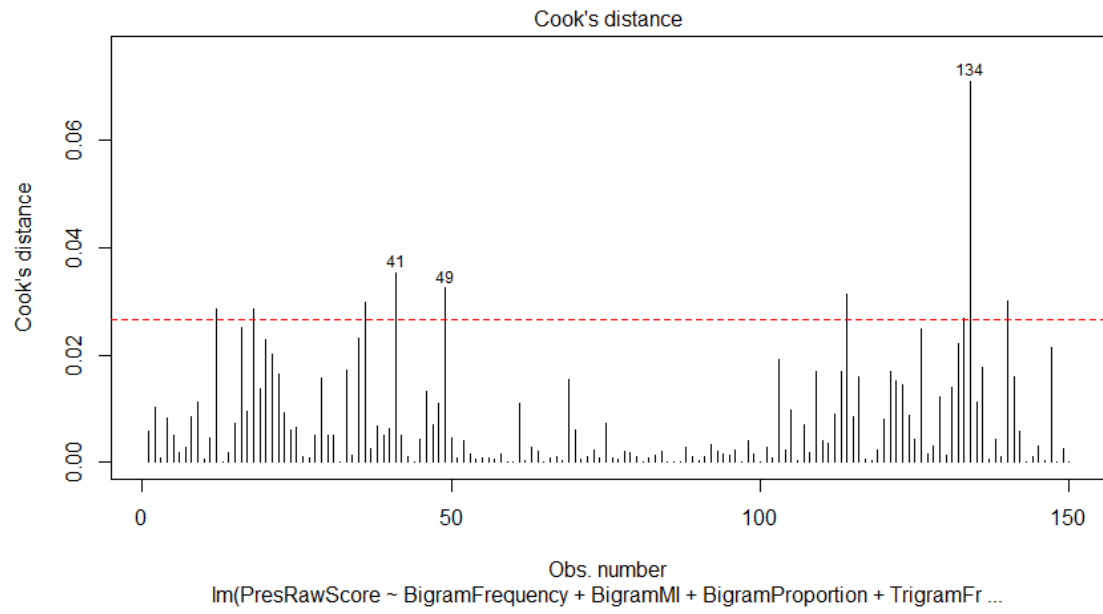
Variance Inflation Factor (VIF) Values

Predictor Variable	VIF Value
Bigram Frequency	2.63
Bigram MI	1.53
Bigram Proportion	4.60
Trigram Frequency	2.08
Trigram MI	1.49
Trigram Proportion	7.21
Three-Five-Word Usage	7.38
Three-Five-Word MI	6.91

Figure S1

Cook's Distance Plot for Detecting Influential Observations in the Regression Model of Presentation Raw Scores

Supplementary Materials



	1X	2Xs	3Xs	4Xs	5Xs		
BigramFrequency		3.80901514	2.907595225	2.909519833	3.5245642735	4.470470709	
BigramProportion		2.73233279	-2.456032629	-6.573170424	-9.7356206673	-11.904197185	
BigramMI		11.29162910	11.871546577	12.428699720	12.9791810099	13.511391696	
TrigramFrequency		3.62051751	2.357859560	1.125215711	-0.0278317499	-1.100005716	
TrigramProportion		18.28787579	18.440385804	17.957124909	16.8607345718	15.107921272	
TrigramMI		0.81550841	0.358790022	-0.168974084	-0.6960828552	-1.170761789	
Three.Five.WordUsage_RawScores		0.25189122	0.283930355	0.323958782	0.3688898582	0.413996918	
Three.Five.WordMI_RawScores		0.01384419	0.009441345	0.004404227	-0.0009271547	-0.006172841	
	6Xs	7Xs	8Xs				
BigramFrequency		5.52236836	6.50784542	7.29418638			
BigramProportion		-12.97046284	-12.84143381	-11.46732976			
BigramMI		13.99119030	14.39475850	14.71759398			

Supplementary Materials

TrigramFrequency	-2.00064757	-2.60771017	-2.80877397
TrigramProportion	12.49047075	8.78610415	3.81287391
TrigramMI	-1.57162558	-1.90155682	-2.18573154
Three.Five.WordUsage_RawScores	0.45559038	0.49251677	0.52527146
Three.Five.WordMI_RawScores	-0.01112559	-0.01576029	-0.02012555

Table S4

R code and detailed results of random forests and Boruta analysis for presentation raw scores

(see corresponding Figure 3 in the main text)

```
library(randomForest)
library(Boruta)
# Random Forest Analysis
```

```
# Random Forest Analysis
```

```
> # Create a random forest model predicting 'PresRawScore' from the MWS variables in the subsetted dataset
```

```
> set.seed(123) # Set the seed for reproducibility
```

```
> forest <- randomForest(PresRawScore~., data=fluencyMWS_subset)
```

```
> print(forest) # Print the summary of the random forest model
```

Call:

```
randomForest(formula = PresRawScore ~ ., data = fluencyMWS_subset)
```

```
  Type of random forest: regression
```

```
  Number of trees: 500
```

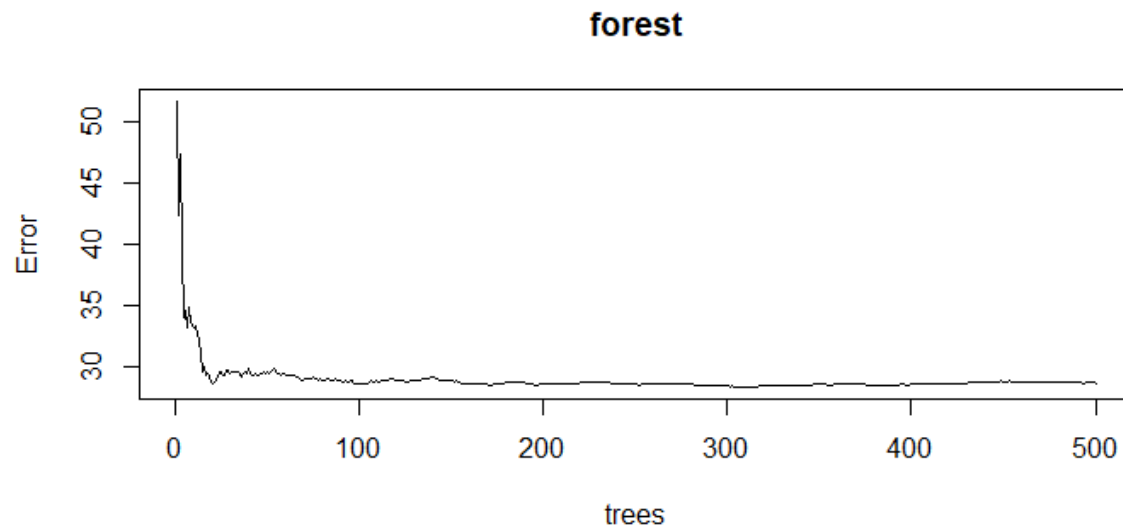
```
No. of variables tried at each split: 2
```

```
  Mean of squared residuals: 28.68067
```

```
  % Var explained: 5.63
```

```
> plot(forest) # Plot the error rates for the random forest model
```

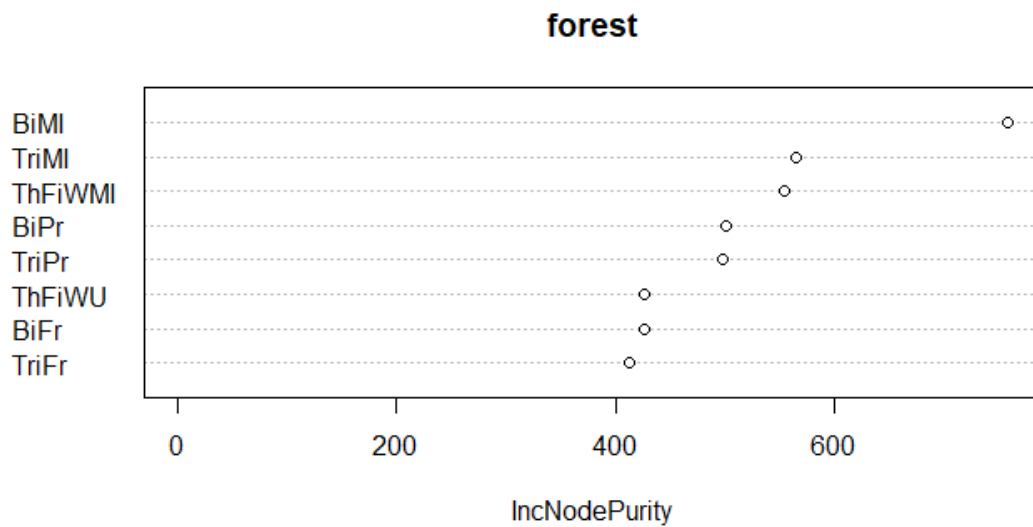
Supplementary Materials



```
> forest$importance # Print the importance of each variable in the model
```

	IncNodePurity
BiFr	426.7300
BiPr	501.1696
BiMI	758.0912
TriFr	413.2714
TriPr	497.2588
TriMI	565.0906
ThFiWU	426.9129
ThFiWMI	553.8998

```
> varImpPlot(forest) # Create a plot showing the importance of each variable
```



Supplementary Materials

> # Boruta Analysis

> print(boruta) # Print the summary of the Boruta results

```
Boruta performed 199 iterations in 4.023549 secs.  
5 attributes confirmed important: BiMI, BiPr, ThFiWMI, ThFiWU, TriPr;  
2 attributes confirmed unimportant: TriFr, TriMI;  
1 tentative attributes left: BiFr;
```

> plot(boruta, cex.axis=0.7, xlab = "Attributes", las = 2) # Plot the Boruta results

> attStats(boruta) # Print detailed attribute (predictor) statistics

	meanImp	medianImp	minImp	maxImp	normHits	decision
BiFr	2.2010065	2.1907797	-0.8444531	5.2091931	0.407035176	Tentative
BiPr	4.9592707	4.9973312	1.0974161	8.1234120	0.849246231	Confirmed
BiMI	9.2754630	9.2674096	5.2658730	13.1527220	0.989949749	Confirmed
TriFr	-1.0054811	-1.0630377	-1.9453361	0.8583341	0.000000000	Rejected
TriPr	5.5457143	5.5469098	2.7947176	8.8312807	0.884422111	Confirmed
TriMI	0.2555087	0.4776789	-1.1449040	1.5994544	0.005025126	Rejected
ThFiWU	8.3270451	8.3700205	5.3404019	12.1870229	0.979899497	Confirmed
ThFiWMI	5.2638682	5.3396385	2.1541636	8.3487941	0.849246231	Confirmed

Environment

[sessionInfo\(\)](#)

```
R version 4.3.0 (2023-04-21 ucrt)  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 11 x64 (build 22631)  
  
Matrix products: default  
  
locale:  
[1] LC_COLLATE=English_United States.utf8 LC_CTYPE=English_United States.utf8  
[3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C  
[5] LC_TIME=English_United States.utf8  
  
time zone: Asia/Tokyo  
tzcode source: internal  
  
attached base packages:  
[1] stats    graphics grDevices utils    datasets methods  base  
  
loaded via a namespace (and not attached):  
[1] compiler_4.3.0 tools_4.3.0  rstudioapi_0.14
```