Korean Journal of English Language and Linguistics, **Vol 25, August 2025, pp. 1125-1149**
DOI: 10.15738/kjell.25..202508.1125

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Can Linguistic Features Predict Placement Decisions? An Exploratory Study on Integrated Summary and Argumentative Writing Samples

Elizabeth Lee (Hankyong National University)

Received: March 29, 2025
Revised: June 10, 2025
Accepted: August 5, 2025

Lee, Elizabeth
Assistant Professor, Department of English Language and Culture, Hankyong National University
327 Jungang-ro, Ansung-si, Gyeonggi-do, Korea 17579
Tel: 031-670-5311
Email: edhlee@hknu.ac.kr

ABSTRACT

Lee, Elizabeth. 2025. Can linguistic features predict placement decisions? An exploratory study on integrated summary and argumentative writing samples. *Korean Journal of English Language and Linguistics* 25, 1125-1149.

This study explores how integrated summary and argumentative essays written by learners at various placement levels can be classified using computational indices measuring cohesion, lexical sophistication, and syntactic complexity. To do so, essays taken from 989 test-takers, whose previous TOEFL scores ranged between 71 and 99, were analyzed using the computational tool Coh-Metrix 3.0. The essays were categorized into B (lowest-level), C, D, and P (highest-level) groups based on raters' decisions. A Random Forest analysis was carried out to predict placement levels using selected Coh-Metrix indices as predictor variables. For summary writing, the top five most important variables predicting an individual's placement level were word familiarity, age of acquisition, word meaningfulness, mean number of words before the main verb, and sentence syntax similarity. For argumentative writing, the five most important predictors were word familiarity, lexical diversity, number of modifiers per noun phrase, word meaningfulness, and age of acquisition. While classification performance was modest overall, the models demonstrated higher precision and recall for P-level essays than for B-, C-, and D-level essays, suggesting that the models experienced difficulty in classifying essays written by students who fall within a narrow proficiency range. Nonetheless, genre-sensitive trends emerged, with argumentative essays showing greater lexical diversity and syntactic elaboration, and summary essays reflecting more gradual increases in sentence complexity. These findings suggest that statistical models may capture linguistic patterns associated with placement levels and offer complementary insights into placement decisions and instructional support.

KEYWORDS

placement test, Coh-Metrix, second language writing, lexical sophistication, syntactic complexity, cohesion, language features, Random Forest

1. Introduction

With the advancement of computational tools, research on predicting writing proficiency and profiles of L1 and L2 students based on linguistic features has expanded. Studies have demonstrated that microlinguistic features can reliably predict or explain proficiency levels (Crossley et al. 2011, Guo et al. 2013), human ratings or test scores (Goh et al. 2020, McNamara et al. 2015), L1 background (Crossley and McNamara 2009, 2016, Crossley and McNamara 2012a), and L2 writing development (Casal and Lee 2019, Crossley et al. 2016). In these bodies of research, studied features typically include those related to cohesion or content overlap (Crossley et al. 2016, Kyle 2020), lexical sophistication (Kyle and Crossley 2016, Vögelin et al. 2019), and syntactic complexity (Casal and Lee 2019, Crossley and McNamara 2014), as well as a combination of two or more of these features (e.g., Crossley and McNamara 2012b, 2016, Ma et al. 2024). These studies have shown that language features not only measure L2 writing development but also predict writing quality and test scores.

Yet, such analyses have yet to be carried out in contexts where locally-developed writing assessments are used to make ESL placement decisions. Unlike large standardized language testing situations (e.g., TOEFL), where the proficiency levels of test-takers are wide-ranging, in locally-developed English placement test contexts, tests are administered to a subset of incoming students who have already met certain admission criteria, and thus the test-taker population's proficiency levels are on a narrower range. An investigation related to these contexts is needed as it can systematically inform whether and to what extent differences in writing quality may vary among students placed into different placement levels. If certain microlinguistic features are found to be significant in predicting group membership, this can shed light on what specific linguistic features ESL instructors should attend to at each level. Another reason for conducting the following research is that in the past, studies analyzing placement test essays have typically focused on a single language feature (e.g., noun phrase complexity, lexical bundles, or lexical frames) and used concordancing as a way to study linguistic differences found across levels (e.g., Jin 2023, Kim 2020, Nguyen 2024, Vo 2019). While there is merit in using such an approach, using a computational tool like Coh-Metrix (McNamara et al. 2014) enables researchers to look at multiple deep-level linguistic features at once and build powerful statistical models for prediction and classification purposes.

A further point to add is that, within assessment contexts, much of the existing work done with Coh-Metrix has examined syntactic complexity, lexical sophistication, or cohesion features within a single genre (e.g., Goh et al. 2020) or on comparisons between two clearly distinct genres, such as independent and integrated writing tasks (e.g., Guo et al. 2013). However, limited attention has been given to related source-based tasks, such as summarizing opposing views followed by writing an argumentative response. In the field of language testing, it is commonly argued that using multiple writing tasks within the targeted domain is crucial for capturing the full range of a learner's writing ability. Given that summarizing and taking a position are both central to academic writing, this has served as a rationale for using source-based summary and argumentative writing tasks in placement testing (Li, 2015). Yet relatively few studies—especially those using computational methods—have examined this notion fully, leaving room for further exploration.

To address the aforementioned research gaps, the current study employs Coh-Metrix (McNamara et al. 2014), a computer-aided automatic text analysis software, to investigate the degree to which lexical sophistication, syntactic complexity, and text cohesion measures could effectively predict students' placement levels using two integrated reading-writing test tasks (i.e., one summary and one argumentative essay). Additionally, a random forest approach is adopted due to its robustness in handling high-dimensional linguistic data and its ability to model complex, non-linear relationships between predictors and outcome categories. The following research questions are posed below, and the next section examines relevant literature pertinent to the focus of this study.

(1) Which linguistic features related to lexical sophistication, syntactic complexity, and cohesion would best classify the integrated summary writing samples according to their placement levels?
(2) Which linguistic features related to lexical sophistication, syntactic complexity, and cohesion would best classify the argumentative writing samples according to their placement levels?
(3) How do these linguistic features differ between the two tasks, and what does this reveal about genre-sensitive placement decisions?

2. Literature Review

2.1 Linguistic Features Used to Predict Placement Decisions

Recent studies have examined linguistic features that can distinguish proficiency levels in writing placement contexts. Vo (2019), for example, focused specifically on lexical features, analyzing both individual words and multi-word sequences in a corpus of 1388 second-language academic placement essays. Using chi-square tests, the author found that higher-proficiency writers produced significantly greater numbers of unique word types, tokens, and word families compared to lower-level writers. However, lexical diversity, density, and sophistication measures did not show clear differences across proficiency groups. Additionally, lower-level essays frequently contained noun-phrase and verb-phrase lexical bundles, whereas advanced-level essays more commonly included prepositional-phrase bundles characteristic of academic writing. These findings suggest that as learners advance in proficiency, their lexical choices shift from more formulaic toward more varied and sophisticated forms.

Jin (2023) similarly investigated lexical patterns, focusing specifically on lexical frames involving the definite article (*the*) in placement test essays by international students at a U.S. university. Analyzing 991 essays classified into lower and higher proficiency groups, chi-square tests were used to identify differences in lexical frame usage. Higher-proficiency essays exhibited more academic-writing frames, greater lexical variability, and less predictable lexical patterns. Conversely, lower-proficiency essays displayed more formulaic and predictable usage, often accompanied by errors involving article misuse or omission. This study highlights the importance of varied and accurate lexical frame use as a marker of proficiency and emphasizes the potential benefits of explicitly teaching these phraseological patterns.

Taking a syntactic perspective, Kim (2020) explored nominal modifiers as predictors of writing placement. Using cumulative ordinal logistic regression with 374 argumentative essays written by international undergraduate students, Kim examined seven types of nominal modifiers previously identified as typical in academic writing. The study showed that four nominal modifiers, namely, prepositional phrases introduced by *of*, premodifying adjectives, multiple prepositional phrases, and prepositional phrases other than *of*, significantly predicted placement decisions. L2 writers who frequently employed these modifiers were more likely to be placed into higher-level courses. Although Kim's study did not employ comprehensive linguistic indices or advanced computational techniques, it clearly demonstrated the predictive value of specific syntactic features in placement contexts.

Finally, Nguyen (2024) also investigated syntactic complexity, focusing specifically on noun phrase complexity in 286 integrated writing placement essays from international students grouped into four proficiency levels (B, C, D, and P). Using one-way ANOVA and non-parametric Kruskal-Wallis tests, the author identified clear proficiency-based differences in noun phrase complexity. Advanced-level writers produced more complex noun modifiers—such as attributive adjectives, concrete and locative prepositional phrases, finite relative clauses

modifying non-animate nouns, and nonfinite relative clauses—while simpler modifiers, such as nouns functioning as premodifiers, did not differ across proficiency groups. Nguyen's findings illustrate a clear developmental progression in syntactic complexity associated with higher proficiency, underscoring the value of detailed grammatical analyses for placement assessment.

In sum, these studies highlight the multifaceted nature of linguistic proficiency in academic writing contexts. Lexical analyses illustrate the role of vocabulary richness and phraseology, while syntactic analyses emphasize structural complexity and grammatical accuracy. Integrating these complementary perspectives through more advanced modeling approaches, such as random forest classification, could further enhance the predictive validity of linguistic features in placement assessments.

2.2 Predicting L2 Proficiency Levels or Ratings Based on Coh-Metrix Indices

In language assessment, writing quality is the extent to which a L2 writer is able to produce texts that align with the goals and purpose of the writing assessment, as judged by human raters (Crossley 2020, Crossley and McNamara 2014). In academic testing contexts, independent and integrated source-based tasks are often used to determine students' academic writing proficiency (Kyle 2020, Guo et al. 2013). Research has shown that highly proficient L2 writers tend to display greater phrasal complexity (Biber and Gray 2010, Casal and Lee 2019), effective use of cohesive devices (McNamara et al. 2015), grammatical (Guo et al. 2013) and sophisticated vocabulary knowledge (Crossley et al. 2011) in their writing than less proficient writers. Good writers have also been found to produce longer texts (Guo et al. 2013, McNamara et al. 2015), and if source texts are present, they tend to engage in greater reproduction of key information, rather than direct copying of source material (Kyle 2020), which would all lead to higher ratings.

Using computational tools like Coh-Metrix, large bodies of research have found positive evidence to support this phenomenon. In terms of lexical sophistication, it has been found that more proficient L2 writers use a wider variety of words as well as low-frequency words in their essays (e.g., Crossley and McNamara 2012b). As well, these have been perceived to be of higher quality by teachers of English (Vögelin et al. 2019). In some studies (e.g., Guo et al. 2013, Kyle and Crossley 2016), proficient writers of English were found to use words having more superordinate terms, as well as fewer polysemous words (e.g., Guo et al. 2013), which would all suggest less ambiguous words being utilized by advanced writers. Sophisticated vocabulary used by more advanced L2 writers, however, has been reported to be less concrete, familiar, imagable, and meaningful (e.g., Crossley et al. 2011, Jung et al. 2019). In other words, advanced writers use words that may be more domain-specific, which suggests their increased knowledge and ability to flexibly use English vocabulary. Words of this nature would also score higher on age-of-acquisition ratings (Ma et al. 2024).

With regards to syntactic complexity using Coh-Metrix, essays receiving higher ratings have been reported to display longer and greater usage of clausal (Crossley 2020) and phrasal complexity (Casal and Lee 2019, Jung et al. 2019). On the other hand, as proficiency advances, adjacent sentence structures become less similar and instead more syntactically varied (Crossley and McNamara 2014). As to cohesion, recent findings suggest that genre may have a strong influence over which cohesive features are deemed appropriate (i.e., given a higher rating) by raters. In Guo et al.'s (2013) study, for example, local cohesion (i.e., semantic similarity between adjacent sentences) was more positively correlated with raters' scores on integrated essays, whereas such a relation could not be found in independent argumentative essays. At the same time, we would also expect a rise in global cohesion as writers become more proficient in their writing (Crossley et al. 2016, Crossley and McNamara 2016, Jung et al. 2019). Based on what has been found in past studies, it would be important to account for genre, which is seen to interact

with linguistic features in complex ways.

In sum, essays written by more proficient L2 writers have been reported to contain more varied words, but those that are less frequently occurring and less easily processed in the mind (Crossley and McNamara 2012b). Depending on the genre of the text, longer clauses and sentences and phrasal elaboration are more likely to be observed in higher-level essays (Casal and Lee 2019). And related to the present study, for source-based writing tasks, we may observe more uses of local cohesive features as well as hypernyms, but lower usage of polysemy, in advanced writers' essays compared to low-level writers, as was found in Guo et al. (2013). Finally, as noted by Crossley and McNamara (2012b) and Kyle (2020), differences in the amount of word and content overlap across levels may also be observed since students had access to source texts during the exam. While previous studies have used individual linguistic features or grouped indices (e.g., Coh-Metrix) to examine proficiency or writing quality, few have employed machine learning algorithms to integrate diverse features for placement-level prediction. This study will explore the degree to which such language features can be observed in essays written in a placement test context using random forest modeling.

3. Method

3.1 Lexical Sophistication

Lexical sophistication refers to the writer's productive knowledge (McNamara et al. 2014), and numerous studies support that lexical sophistication and writing quality are positively related (e.g., Kyle and Crossley 2016, Vögelin et al. 2019). Lexical sophistication is a multidimensional construct that can be evaluated in a number of different ways in Coh-Metrix, which include measuring lexical diversity, word frequency, semantic relationships (i.e., hypernymy and polysemy), and psycholinguistic properties of words (i.e., concreteness, familiarity, imagability, meaningfulness, and age-of-acquisition scores) (McNamara et al. 2014). In this study, lexical diversity (LDMTLD) was measured using a special index known as the Measure of Textual Lexical Diversity that is applied to all words. A high value in lexical diversity would indicate more unique words being used in the text. For word frequency (WRDFQmc), the average minimum log word frequency for content words that could be found in the CELEX corpus was estimated. A large average value indicates more high-frequency content words being present in a given text. The semantic dimension of words was calculated using the average WordNet hypernymy (WRDHYPnv) and polysemy (WRDPOLc) indices. WRDHYPnv measures how concrete (non-abstract) the nouns and verbs appear to be, whereas WRDPOLc measures the degree to which words appear ambiguous. A high score on hypernymy and a low score on polysemy would generally suggest writers using more concrete, non-abstract words.

The psychological dimension of words is estimated using five indices that are drawn from the MRC Psycholinguistic Database: concreteness (WRDCNCc), familiarity (WRDFAMc), imagability (WRDIMGc), meaningfulness (WRDMEAc), and age of acquisition (WRDAOAc). Concreteness refers to how concrete or non-abstract a word appears, as judged based on human ratings. Familiarity is the extent to which a word appears familiar to an adult due to ease of processability. Imagability points to how easily a mental image can be constructed given a word. Meaningfulness refers to the extent to which a word is associated with other words. And last, age of acquisition is an index of ratings that indicate how early (or late) a spoken word is thought to be acquired by children. In summary, higher ratings on the first four MRC-based indices would suggest that a writer is using more concrete, familiar, imagable, and strongly associative words. A higher rating on age of acquisition,

however, would suggest writing using words that are assumed to be acquired later in life.

3.2 Syntactic Complexity

Syntactic complexity is defined as the degree to which a writer is able to use a wide range of sophisticated grammatical structures (McNamara et al. 2014). Similar to lexical sophistication, writing quality and proficiency tend to rise with increased display in syntactic complexity (e.g., Casal and Lee 2019, Ma et al. 2024). Coh-Metrix conducts more fine-grained syntactic complexity analysis, and in this study, three different measures were considered: mean number of words before the main verb (SYNLE), mean number of modifiers per noun phrase (SYNNP), and syntactic similarity (SYNSTRUTt). SYNLE and SYNNP are indices that estimate how syntactically dense sentences appear in a text. The more words appear before the main verb or modifiers before a head noun, the more syntactically complex and difficult it is to process efficiently. In academic writing, information tends to get structurally compressed (Biber and Gray 2010, Casal and Lee 2019), and so we would expect to observe higher estimates of SYNLE and SYNNP. SYNSTRUTt, on the other hand, measures how syntactically similar adjacent sentences are to each other. The more syntactic similarity we observe, the greater the text easability. High syntactic similarity, though making the text easier to read, would suggest a lack of sentence variety in the text, yet using a range of sentence structures is an important element in academic writing. An additional three variables, the relative density of prepositional phrases (DRPP), noun phrases (DRNP), and verb phrases (DRVP), were also included. According to McNamara et al. (2014), an increase in the relative density of each feature would raise difficulty in text processing, as such phrases tend to convey a dense amount of information through complex sentence structures. In academic writing, L2 writers displaying higher levels of phrasal complexity have been shown to earn higher instructor ratings of writing quality (Casal and Lee 2019), and so these features were examined in the present study.

3.3 Cohesion

Cohesion in Coh-Metrix refers to how different elements in a text connect to enhance readability and comprehension (McNamara et al. 2014). Crossley et al. (2016) and other similar studies have found evidence that writing development was positively associated with greater local and global text cohesion. Cohesive devices operating at the sentence level are considered local, such as the use of pronouns, conjunctions, and synonyms in one's writing. These devices serve to ensure that sentence-to-sentence ideas connect smoothly. Those operating across paragraphs or larger sections in a text are considered global. For example, topic sentences at the start of each paragraph and content word overlap occurring over larger stretches of text are intended to create greater uniformity and clarity in one's writing. Both types of cohesive devices were explored in the current study, where lexical co-referentiality (CRFSOl), semantic co-referentiality (LSASSl), minimal edit distances (SYNMEDlem), causality (SMCAUSr), temporal markers (CNCTemp), logical operators (CNCLogic), and lexical diversity (LDMTLD) estimate the presence of local cohesion; and LSA givenness (LSAGN), WordNet verb overlap (SMCAUSwn), and temporal cohesion (SMTEMP) estimate the presence of global cohesion. In Crossley et al.'s (2016) study, lexical diversity was also used as a measure of cohesion, and so this approach was adopted.

To briefly describe each index, CRFSOl measures the repetition of content words between sentences. Higher scores indicate greater word overlap, which facilitates the readability of a text. LSASSl extends this idea by examining the overlap of words that are similar in meaning, such as "home" and "house," found in adjacent sentences. A higher score on LSASSl indicates greater semantic overlap which creates better text cohesion. Other

cohesion measures include SYNMEDlem, which gauges sentence variation—higher values indicate greater structural differences. SMCAUSr evaluates the presence of causal connections (e.g., "because," "as a result") in writing, with higher scores suggesting stronger links between actions and events. Cohesion is further reinforced through connectives. CNCTemp ("first," "until") and CNCLogic ("and," "or") contribute to a more structured and cohesive text. And a high value in LDMTLD, in terms of cohesion, would suggest that a text demonstrates less cohesiveness.

SMCAUSwn measures verb overlap using a WordNet algorithm, which indicates how often a verb is repeated across a text. Verb repetition tends to decrease as texts become more complex. LSA givenness, LSAGN, is another kind of semantic co-referentiality measure that assesses the ratio of new versus previously mentioned information in a text. A higher score on LSAGN implies that a writer has provided more given than new information, which is said to improve text cohesion. SMTEMP measures temporal cohesion, specifically the number of times the tense and aspect of the main verb or helping verb are repeated across a text. The repetition score too decreases the more times temporal shifts occur. In sum, 21 linguistic features related to lexical sophistication, syntactic complexity, and cohesion were used to conduct the present study.

3.4 Sample Selection

To answer the research question, Iowa State University's (ISU) English Placement Test (EPT) Corpus of Learner Writing 2.2 (2017) was used (N = 991 test takers). This corpus consisted of 991 summary essays and 991 argumentative essays drawn from the writing portion of the English placement test that was used for placement purposes at a large midwestern university between 2016 and 2017. The essays were handwritten by undergraduate- and graduate-level international students whose English was not their native language. Students were given 50 minutes to complete two integrated writing tasks: (1) a summary essay of 100-150 words summarizing two reading passages, and (2) an argumentative essay of 300-350 words defending one's position on a given topic related to the two reading texts. Between 2016 and 2017, at least three different topics were used. Although topic information and reading passages were not accessible, it is assumed that the difficulty of test topics and reading passages was relatively similar. For the corpus, the essays were transcribed, making no corrections to grammar or spelling; undecipherable words were marked with brackets.

Placement decisions were made by two to three raters, and each student received one of four placement levels: B, C, D, or P. Students assigned to level B, the lowest tier, participated in a semester-long writing course emphasizing English grammar and paragraph-level composition. Those in level C were undergraduate students who received instruction in basic academic writing skills. Meanwhile, students placed in level D were graduate students who focused on research writing. And students earning a P were exempted from taking any remedial ESL courses. Details about the corpus can be seen in Table 1. It should be noted that due to some formatting differences, essays written by two test-takers were excluded, and as a result, 989 sets of summary and argumentative essays were sampled.

Table 1. Description of the Original Corpus (Release 2.2. 2017, adapted from Danis 2019)

		Task Type	
Level		Summary	Argumentative Writing
B	Number of texts	202	202
	Number of words	23,771	40,881
	Words per text	117.68	202.38
C	Number of texts	298	298
	Number of words	37,066	69,225
	Words per text	124.38	292.30
D	Number of texts	157	157
	Number of words	21,825	37,201
	Words per text	139.01	236.95
P	Number of texts	334	334
	Number of words	46,343	86,687
	Words per text	138.75	298.54
Total number of words in corpus		363,341	

Because this study used both the summary and argumentative writing essays, the original corpus was divided into two sub-corpora, and separate analyses were conducted on each sub-corpus. All meta-tags were removed, and no corrections were made before loading the texts to Coh-Metrix 3.0. The output, which is generated in a .csv format, was uploaded to R 4.5.0 (R Core Team 2025) to conduct the data analysis.

3.5 Statistical Analyses

In this study, 21 variables related to cohesion, lexical sophistication, and syntactic complexity were initially selected from the Coh-Metrix output. These measures were well-validated in past studies (see Crossley 2020, McNamara et al. 2014) and were reported to be useful for determining differences among speakers of English at different proficiency levels (Crossley and McNamara 2012a). Before running any statistical tests, each sub-corpus was divided into an 80/20 split using R's *set.seed()*: a training set (790 essays) and a test set (199 essays) (see Table 2). This division was necessary to ensure that the generated model was reliable and generalizable. The data sets were stratified by placement level (B, C, D, P) and treated as an ordered factor to reflect increasing proficiency.

Table 2. Description of Class Size for Training and Test Sets

Level	Training set	Test set
B	160 (20.3%)	40 (20.1%)
C	238 (30.1%)	60 (30.2%)
D	125 (15.8%)	32 (16.1%)
P	267 (33.8%)	67 (33.7%)
Total	790	199

Because discriminant function analysis (DFA) has been a common approach to distinguishing writing quality (e.g., McNamara et al. 2015) and texts written by L1 and L2 writers of English (e.g., Crossley and McNamara 2009), initially, this study adopted a quadratic discriminant analysis (QDA), a statistical classification method, as a means of exploring whether placement decisions could be predicted based on Coh-Metrix variables. However, because QDA assumes a normal distribution of data, and this assumption was not met, results generated from the QDA analysis were weak and unreliable. Consequently, Random Forest Classification (RFC), a tree-based ensemble learning method that does not assume normality, was used to overcome normality violations and class

imbalance. Additionally, RFC was selected because it is generally robust to multicollinearity and can handle a large number of predictor variables without overfitting or requiring transformation of the input data. What is more, RFC generates each variable's importance in prediction, which is important in investigating which linguistic features are most significant in predicting placement levels. While uses of Random Forest (RF) methods are not as widely observed as DFA for classification purposes, a small but emerging number of studies have shown that RF can model complex, multidimensional language data (e.g., Mizumoto 2023, Zagata et al. 2023).

All analyses were conducted in R using the packages *tidymodels*, *themis*, *ranger*, *vip*, *pdp*, and *yardstick*, which support model tuning, evaluation, and visualization. After dividing the data into training and test sets, a random forest classifier model was implemented with only the training set. In the dataset, the D-level placement group contained fewer essays than the other placement levels. To address this class imbalance, a hybrid resampling technique known as Synthetic Minority Oversampling Technique (SMOTE) combined with Tomek links (Batista et al. 2004) was introduced. The SMOTE algorithm creates synthetic essays for the minority classes (in this case, level D) that are similar to existing samples in the dataset, thereby increasing the presence of D-level essays in the RF classifier model. Next, the Tomek links method, an undersampling technique, was applied to further clean the training data. For example, if a D-level essay is highly similar in its linguistic features to a neighboring C-level or P-level essay, this pair would be flagged as a Tomek Link and thus removed. In theory, the removal of such borderline cases would make it easier for the RF classifier model to learn different placement categories, especially where essays from different levels overlap in their linguistic features. All numeric predictors were normalized, and predictors with no variation (zero-valence predictors) were removed prior to modeling.

To ensure robust model evaluation and tuning, a special hyperparameter tuning known as nested 5 x 5 nested cross-validation (CV) was employed, with macro-averaged F1 score as the tuning objective. This is an approach that trains and tests the model on different parts of the data to avoid overfitting. Macro-F1 calculates the F1 score for each class independently and then averages them, giving equal weight to each placement level. This ensures that the model performance is evaluated fairly across both majority and minority classes, and is particularly appropriate for imbalanced multiclass classification as seen in the present study. The hyperparameter tuning used grid search, a method that tests different values to see which model settings work best. Specifically, it evaluated 25 unique combinations of the number of predictors sampled at each split (mtry) and minimum node size (min_n). The final model was fitted using the optimal hyperparameters derived from the nested CV process, and its performance was evaluated on the held-out test set using overall accuracy, adjacent accuracy, and macro-averaged precision, recall, and F1 score. Finally, variable importance was evaluated using mean decrease in Gini index (a measure of how each feature reduces classification error), and the top 10 most influential features were plotted. Additionally, partial dependence plots (PDPs), which show the predicted effect of one variable while keeping other variables constant, for the top five most influential features across all four placement levels, were generated to further examine their contribution to placement classification.

4. Results

4.1 Summary Essay Set

4.1.1 Model performance

The final random forest model was built with 1000 trees, and hyperparameter tuning via 5x5 nested CV selected

an optimal configuration with an mtry value (i.e., the number of predictors randomly sampled at each split) of 12 and a min_n value (i.e., tree depth) of 6. Across the five outer folds, the model achieved a mean accuracy of 0.379, a Cohen's kappa of 0.151, and a macro-averaged F1 score of 0.342 (SD = 0.034), indicating only a modest but stable classification performance in differentiating placement levels based on linguistic features. This result reflects that even after SMOTE-Tomek was applied, the model had difficulty in distinguishing adjacent proficiency levels using automated linguistic measures alone. On the held-out test set (20% of the data), the model achieved a macro-averaged F1 score of 0.324, an overall classification accuracy of 0.342, an adjacent accuracy rate of 0.683, and a Cohen's kappa of 0.089. Macro precision and recall values were 0.328 and 0.323, respectively. These numbers suggest that the model was only occasionally able to approximate placement decisions within one level of the true class, and that the overall precision in identifying the exact placement level remained strongly limited.

Taken as a whole, these results indicate that while the model captured some general trends in linguistic differences across placement levels, especially for P-level essays, classification was less consistent among the lower and adjacent levels (i.e., B, C, and D). The relatively low macro-F1 scores likely reflect both the difficulty of distinguishing writing placement levels based on automated linguistic features and that there was in fact more overlap in linguistic features found across different placement levels.

4.1.2 Variable importance

Figure 1 shows the top 10 most influential predictors contributing to placement classification. The top-ranked features, based on mean decrease in Gini index, included measures of lexical sophistication, syntactic complexity, and cohesion. Specifically, WRDFAMc (51.3) emerged as the most influential predictor, followed by WRDAOAc (39.0) and WRDMEAc (35.1). Several syntactic complexity measures also ranked fairly highly, including SYNLE (34.0), SYNSTRUTt (33.3), and SYNNP (32.4). Cohesion measures such as LSAGN (31.9), SYNMEDlem (31.8), SMCAUSwn (31.3), and LDMTLD (30.8), were among the top predictors as well. In Table 3, descriptive statistics for the top 10 most important linguistic features were calculated using the original dataset prior to preprocessing.

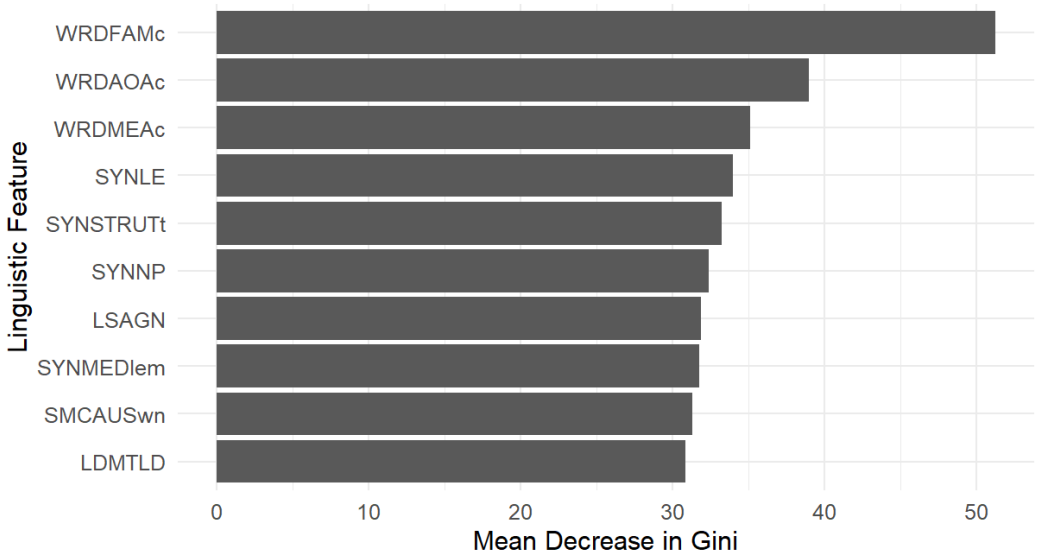

Figure 1. Top 10 Features for Placement Prediction – Summary Essays

Table 3. Descriptive Statistics for Top 10 Linguistic Features from Summary Corpus

Variables	Importance	B	C	D	P
WRDFAMc	51.3	578.526	576.531	572.981	571.545
		(7.729)	(8.859)	(8.811)	(7.927)
WRDAOAc	39.0	377.681	385.815	387.792	390.447
		(30.245)	(30.234)	(23.728)	(23.451)
WRDMEAc	35.1	429.023	425.936	427.200	422.853
		(18.734)	(18.748)	(13.619)	(18.366)
SYNLE	34.0	4.279	4.430	5.040	4.641
		(2.426)	(1.818)	(2.562)	(2.143)
SYNSTRUTt	33.3	0.105	0.101	0.091	0.092
		(0.039)	(0.032)	(0.030)	(0.034)
SYNNP	32.4	0.742	0.738	0.807	0.806
		(0.213)	(0.173)	(0.169)	(0.169)
LSAGN	31.9	0.265	0.254	0.268	0.253
		(0.070)	(0.060)	(0.051)	(0.055)
SYNMEDlem	31.8	0.846	0.853	0.856	0.851
		(0.101)	(0.069)	(0.038)	(0.068)
SMCAUSwn	31.3	0.499	0.510	0.493	0.504
		(0.115)	(0.120)	(0.110)	(0.118)
LDMTLD	30.8	75.795	79.357	84.307	83.105
		(24.413)	(23.542)	(21.388)	(23.504)

These findings suggest that a combination of measures related to lexical sophistication, syntactic complexity, and cohesion contributed to the model's classification decisions, especially at higher placement levels. However, it is important to note that variable importance values in random forests indicate which features were most useful for separating essays into placement levels within the model, but they do not necessarily imply direct causal relationships between specific linguistic features and placement outcomes. For example, WRDFAMc emerged as the most influential feature, suggesting that essays containing less familiar vocabulary tended to be associated with higher placement levels in the dataset. However, this does not necessarily mean that decreasing word familiarity alone would directly cause a student to be assigned to a higher level. To clarify, essays at higher levels tended to contain vocabulary that was, on average, less familiar. This pattern suggests that lower word familiarity often co-occurred with more advanced writing, though other unmeasured factors likely also contributed to placement decisions.

4.1.3 Partial dependence analyses

To further examine how key predictors influenced placement probabilities, partial dependence plots (PDPs) were generated only for the top five features across all four placement levels (see Appendix A). For WRDFAMc, increasing WRDFAMc values were associated with a higher probability of a B- or C-level placement, while the probability of a D- or P-level placement decreased beyond the mean-centered value of 0. For example, when WRDFAMc increased above 0, the probability for P-level essays notably declined from approximately 0.35 to below 0.20. Similarly, WRDAOAc showed that lower age-of-acquisition scores (i.e., using less advanced, more commonly acquired words) increased the probability of B-level placement while decreasing the probability of P-level placement. However, some irregularities were observed for C- and D-level predictions associated with the WRDAOAc variable, potentially reflecting data sparsity issues and interactions with other linguistic features. On

the other hand, the variable WRDMEAc revealed a nonlinear relationship, where WRDMEAc values around 0 were associated with increased probabilities for D-level placement, while both extremely low and high values reduced the likelihood of being classified into this level. Conversely, lower WRDMEAc values increased the probability of P-level classification, suggesting that higher-level writers tended to draw on more abstract, academic language.

For syntactic complexity, both SYNSTRUTt and SYNLE exhibited nonlinear association with placement levels, where predicted probabilities shifted as the feature values crossed certain points. Increasing SYNSTRUTt was generally associated with higher predicted probability for B- and C-level placements, while higher SYNSTRUTt values corresponded to lower probability for D- and P-level essays. For example, when SYNSTRUTt exceeded approximately 1.5, the probability of D-level placement dropped sharply, suggesting that greater sentence variety (i.e., a lower SYNSTRUTt value) is associated with higher placement levels (D or P). For SYNLE, the PDP showed that a higher mean number of words before the main verb was positively related to D-level placement but negatively associated with B, C, and P levels. While it might be theoretically expected that P-level essays would also demonstrate even higher levels of left-embedding, this pattern was not clearly observed in the model outputs, possibly due to structural variation beyond what SYNLE captures or due to limited sample size. On the one hand, it could be that P-level essays employed a wider variety of complex structures beyond left-embedding that may not be fully captured by SYNLE alone. On the other hand, the relatively small sample size may have limited the model's ability to detect usage of left-embeddedness for P-level writers. Taken together, these partial dependence patterns highlight nuanced relationships between linguistic features and placement levels, underscoring that the same feature may exert different influences depending on which level is being predicted. Nevertheless, these effects should be interpreted cautiously as PDPs reflect model-estimated associations and do not imply direct causality.

4.1.4 Classification confusion matrix patterns

The confusion matrix (Table 4) revealed that misclassifications were common, tending to concentrate around adjacent levels, suggesting that most errors involved neighboring proficiency levels rather than entirely unrelated placements. For B-level essays, only 8 were correctly classified, and the most frequent misclassification was into C-level (15 essays) or P-level (10 essays). Similarly, for C-level essays, only 18 were correctly classified, but they were often confused with both B (20 essays) and P level (14 essays). For D-level essays, a substantial portion of the essays was assigned to the P (11 essays) or C level (4 essays), but only 10 were correctly classified. P-level essays exhibited the highest correct classification rate (32 essays), though a good number of them were still predicted as C-level (23 essays). These patterns indicate that the model's errors largely involved difficulty in making fine-grained distinctions between adjacent placement levels. Finally, Cohen's kappa was 0.089 (multiclass), indicating that while the model was able to capture some systematic patterns in the data, its ability to make reliable placement distinctions remained low. This reflects some degree of underlying linguistic continuity across placement levels, and that distinguishing adjacent placement levels using limited features can be challenging.

Table 4. Predicted Versus Actual Text Type for the Summary Corpus

Predicted text type	Actual text type			
Test set	B	C	D	P
B	**8**	15	4	10
C	20	**18**	6	14
D	2	4	**10**	11
P	10	23	12	**32**

Table 5. Precision, Recall, F1, and Adjacent Accuracy for the Summary Corpus

Test set	Precision	Recall	F1	Adjacent accuracy
B	0.216	0.200	0.208	0.700
C	0.310	0.300	0.305	0.617
D	0.370	0.312	0.339	0.875
P	0.416	0.478	0.444	0.642

Table 5 reports per-class precision, recall, F1 scores, and adjacent accuracy of the models predicting placement decisions based on the summary essays. For B-level essays, both precision (0.216) and recall (0.200) were low, resulting in an F1 score of 0.208, suggesting substantial difficulty in accurately identifying lower-level essays. Performance was somewhat better for C- and D-level essays with precision increasing to 0.310 and 0.370, and corresponding recall scores moving to 0.300 and 0.312. The highest performance was observed for P-level essays, where precision reached 0.416 and recall 0.478, resulting in the highest F1 score (0.444) among the four levels. Adjacent accuracy, on the other hand, was highest for D-level essays (0.875) than for C-level essays (0.617), despite the two levels being neighboring categories. Whereas misclassifications for D-level essays tended to remain within the immediately adjacent categories (C or P), C-level essays were misclassified into both lower (B) and non-adjacent higher levels (P), suggesting that C-level students may be showing greater heterogeneity in their writing performance, as would be expected of intermediate-level writers.

4.2 Argumentative Writing Set

4.2.1 Model performance

Next, we turn to the results found from the argumentative writing corpus. Using a total of 1000 trees, the final RF model was trained, with hyperparameter tuning performed through a 5x5 nested cross-validation. This process yielded an optimal set of parameters: mtry at 10 and min_n at 6. Across the five outer folds, the model produced a mean accuracy of 0.377, a Cohen's kappa of 0.145, and a macro-averaged F1 score of 0.343 (SD = 0.013), which are generally consistent with the pattern observed for the summary essay dataset. Even with the application of SMOTE-Tokek resampling to address class imbalance, the model continued to exhibit limited capacity to reliably separate adjacent placement levels solely based on automated linguistic features.

Performance on the held-out test set, comprising 20% of the full sample, showed similar tendencies: a macro-averaged F1 score of 0.397, an overall classification accuracy of 0.167, an adjacent accuracy rate of 0.739, and a Cohen's kappa of 0.365. As in the summary dataset, F1 and kappa values suggest that while some patterns were learned by the model, exact classification remained difficult. Moreover, macro precision and recall values were 0.367 and 0.366, respectively. Although the model's performance was slightly better than the model done on the summary dataset, its precision in exact classification remained limited. These findings suggest that while the model was able to detect certain broad patterns of linguistic variation across placement levels—particularly for P-level

essays—the distinction among the lower and neighboring levels remained less consistent. As was observed with the summary essays, the relatively low macro-F1 values reveal the model's difficulty in predicting placement levels using linguistic features alone, and that there exists considerable overlap in linguistic characteristics across adjacent levels.

4.2.2 Variable importance

Figure 2 shows the top 10 most influential predictors contributing to placement classification using the argumentative writing dataset. The top-ranked features, based on mean decrease in Gini index, again suggest that measures of lexical sophistication, syntactic complexity, and cohesion play a strong role in argumentative writing classification; however, the order of importance somewhat varied. WRDFAMc (47.8) emerged as the most influential predictor, followed by LDMTLD (36.4) and SYNNP (35.9). WRDMEAc (33.1), WRDAOAc (33.0), and DRPP (32.4) ranked next. Finally, WRDFRQmc (30.5), CNCLogic (30.1), SMTEMP (30.0), and SYNSTRUTt (29.8) followed afterward. Table 6 shows the descriptive statistics for these top 10 most important linguistic features that were calculated using the original dataset.

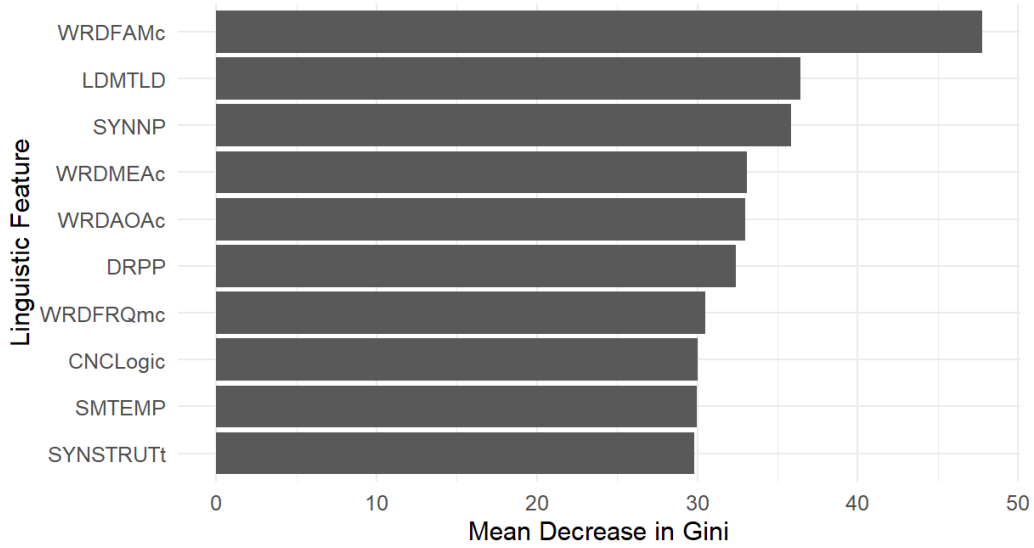

Figure 2. Top 10 Features for Placement Prediction – Argumentative Writing Essays

These results suggest that linguistic measures reflecting lexical sophistication (WRDFAMc, WRDMEAc, WRDAOAc, WRDFRQmc), syntactic complexity (SYNNP, DRPP, SYNSTRUTt), and cohesion (LDMTLD, CNCLogic, SMTEMP) contributed to the model's ability to differentiate placement levels in argumentative writing. At the same time, it is important to recognize that, as seen with the summary RF model, variable importance rankings in random forests reflect the relative usefulness of features for classification within the dataset, rather than indicating direct causal relationships.

Table 6. Descriptive Statistics for Top 10 Linguistic Features from Argumentative Corpus

Variables	Importance	B	C	D	P
WRDFAMc	47.8	581.326	580.698	577.568	576.549
		(6.272)	(6.743)	(6.179)	(6.266)
LDMTLD	36.4	76.144	75.953	87.389	84.547
		(19.067)	(17.713)	(20.593)	(17.891)
SYNNP	35.9	0.651	0.651	0.723	0.723
		(0.170)	(0.146)	(0.150)	(0.154)
WRDMEAc	33.1	427.866	426.695	426.920	422.231
		(16.383)	(15.984)	(15.552)	(14.988)
WRDAOAc	33.0	358.088	359.765	369.771	371.230
		(32.393)	(30.272)	(26.892)	(26.509)
DRPP	32.4	96.077	100.307	108.952	108.785
		(24.784)	(23.079)	(22.875)	(22.658)
WRDFRQmc	30.5	1.217	1.206	1.115	1.055
		(0.346)	(0.330)	(0.313)	(0.320)
CNCLogic	30.1	54.919	53.581	48.415	49.534
		(19.329)	(18.668)	(16.048)	(17.025)
SMTEMP	30.0	0.783	0.782	0.781	0.756
		(0.222)	(0.094)	(0.389)	(0.100)
SYNSTRUTt	29.8	0.105	0.103	0.098	0.090
		(0.032)	(0.030)	(0.025)	(0.025)

4.2.3 Partial dependence analyses

Next, partial dependence plots (see Appendix B) were examined for the top five most influential predictors to illustrate how individual linguistic features were associated with placement level classifications using the argumentative writing dataset. For WRDFAMc, higher scores were associated with increased probabilities of being classified into C and B levels, while probabilities for P and D declined, particularly beyond the standardized mean. For LDMTLD, greater lexical diversity increased D- and P-level classification, whereas decreasing observations in B and C levels, suggesting that a stronger association between lexical range and upper placement performance exists. As for SYNNP, the plot shows that a higher mean number of words per noun phrase was linked with increased probabilities for D- and P-level placement, while those probabilities for B and C decreased. For WRDMEAc, higher word meaningfulness scores tended to favor B and D classification, with lower probabilities for P and C, suggesting that more abstract word use may somewhat characterize higher placement levels. Finally, higher WRDAOAc scores were more positively associated with D, whereas probabilities for B and P diminished or demonstrated little change across the range of WRDAOAc scores.

The partial dependence trends revealed that many of these linguistic features distinguished between the lower (B and C) and higher (D and P) placement levels, but did not necessarily follow a uniform upward trend from B to P. Higher scores on WRDFAMc and WRDMEAc tended to be observed among students placed at lower-to-mid levels (B and C), whereas higher scores on LDMTLD, SYNNP, and WRDAOAc were more frequently associated with graduate students placed at upper levels (D and P), reflecting greater lexical diversity, syntactic complexity, and advanced vocabulary usage.

4.2.4 Classification confusion matrix patterns

The confusion matrix (Table 7) indicated that classification errors were likewise frequent and that most

misclassifications occurred between neighboring proficiency categories. For B-level argumentative essays, 12 were correctly classified, while most errors involved assignment to C-level (12 essays) or P-level (8 essays). C-level essays showed a similar pattern, with 20 correctly classified and frequent misclassifications into B-level (17) and P-level (13). For D-level essays, only 8 were correctly classified, while misclassified essays were assigned to C-level (11) or P-level (7). P-level essays had once again the highest number of correct classifications (39), though misclassification into C-level (17) and D-level (15) was still common. These patterns suggest that the model struggled to make fine distinctions between closely related placement levels in argumentative writing. The overall Cohen's kappa was 0.167, reflecting minimal agreement beyond chance. While the model was able to capture some distributional patterns, its ability to assign essays to exact placement levels was hardly better than what was observed in the summary essay dataset.

Table 7. Predicted Versus Actual Text Type for the Argumentative Writing Corpus

Predicted text type	Actual text type			
Test set	B	C	D	P
B	**12**	12	3	8
C	17	**20**	6	13
D	4	11	**8**	7
P	7	17	15	**39**

Table 8. Precision, Recall, F1, and Adjacent Accuracy for the Argumentative Writing Corpus

Test set	Precision	Recall	F1	Adjacent accuracy
B	0.343	0.300	0.320	0.725
C	0.357	0.333	0.345	0.717
D	0.267	0.250	0.258	0.906
P	0.500	0.582	0.538	0.687

Table 8 summarizes the per-class precision, recall, F1 scores, and adjacent accuracy for the model predicting placement levels based on the argumentative essays. For B-level essays, precision was 0.343 and recall was 0.300, yielding an F1 score of 0.320, indicating that accurately identifying essays at the lower end of the placement scale remained challenging. Performance for C-level essays showed slight improvement, with precision at 0.357, recall at 0.333, and an F1 score of 0.345. For D-level essays, the model achieved a precision of 0.267, a recall of 0.250, and the lowest F1 score of 0.258 among all levels. In contrast, classification of P-level essays showed stronger results, with precision rising to 0.500 and recall to 0.582, and the highest F1 score of 0.538.

Adjacent accuracy followed a somewhat different pattern. The highest adjacent accuracy was observed for D-level essays (0.906), suggesting that although exact classification was limited, the model tended to assign these essays to immediately neighboring levels. C-level essays, in comparison, exhibited an adjacent accuracy of 0.717, reflecting greater dispersion of misclassifications into both lower (B) and non-adjacent higher (P) levels. Given that wider variability in linguistic performance typically persists with intermediate-level writers, it is not unexpected that such a distribution was observed among C-level students.

5. Discussion

5.1 Most Important Linguistic Features for Placement Predictions from the Summary Corpus

Consistent with earlier studies employing Coh-Metrix indices (Crossley et al. 2011, McNamara et al. 2014, Jung et al. 2019), measures of lexical sophistication, including word familiarity (WRDFAMc) and age of acquisition (WRDAOAc), were associated with placement level distinctions. Essays written by higher-placed students tended to contain less familiar and later-acquired vocabulary, reflecting patterns of lexical development observed in prior research. Similarly, word meaningfulness (WRDMEAc) differentiated placement groups to a degree, with lower-level writers using more highly meaningful and easily processed words, while higher-level writers incorporated more abstract and less predictable vocabulary (Crossley et al. 2011, McNamara et al. 2014).

Syntactic complexity measures revealed comparable trends as well. Essays assigned to higher placement levels exhibited greater use of left-embedded structures (SYNLE) and less sentence similarity (SYNSTRUTt). This aligns with previous findings that more proficient academic writers produce syntactically denser and more structurally diverse texts (Biber and Gray 2010, Crossley and McNamara 2014, Goh et al. 2020). On the other hand, cohesion measures were ranked less important than lexical sophistication or syntactic complexity measures, suggesting that while cohesion is important for text comprehensibility, it may be less strongly associated with placement level differences. Although this finding contrasts with previous studies such as Guo et al. (2013), which reported positive associations between integrated essays and cohesion, it may be that in the present study, L2 writers within a relatively narrow proficiency band use similar cohesive strategies. This would make it difficult to detect significant differences across levels. Alternatively, it may be that since source texts were made available during the test, variation in cohesion among the four placement levels could not be detected as strongly as in other studies (e.g., Jung et al. 2019). Finally, the model's ability to classify essays with high precision was modest, with substantial overlap across adjacent placement levels. These findings suggest that while linguistic features can capture broad trends in placement decisions, as noted in previous research (e.g., Kim 2020), they cannot fully capture the nuanced judgments often involved in human rating.

5.2 Most Important Linguistic Features for Placement Predictions from the Argumentative Corpus

The analysis of the argumentative writing dataset offers further insight into how linguistic features contribute to placement classification in L2 writing. Word familiarity (WRDFAMc) once again emerged as a key feature, with lower familiarity scores more frequently observed in essays assigned to higher placement levels, reflecting patterns of increasing lexical sophistication reported in previous research (Crossley et al. 2011, Jung et al. 2019). Lexical diversity (LDMTLD) also differentiated placement levels, with higher values associated with D- and P-level essays, in concert with findings that advanced L2 writers produce texts containing a broader range of unique vocabulary (Vögelin et al. 2019). At the same time, this may also reflect reduced cohesion commonly observed during L2 development (Crossley and McNamara 2012b, 2014).

Syntactic complexity, as indexed by the number of modifiers per noun phrase (SYNNP), likewise distinguished placement levels, with higher values found in essays produced by higher-placed writers. This aligns with prior studies highlighting the role of complex noun phrases in advanced academic writing (Biber and Gray 2010, Casal and Lee 2019, Jung et al. 2019). Word meaningfulness (WRDMEAc) and age of acquisition (WRDAOAc) also contributed to the model's classifications, though their patterns were more variable across placement levels. Nonetheless, essays written by higher-level students did include a greater number of later-acquired words as well as words that are less easy to process (McNamara et al. 2014). And finally, polysemy and hypernymy were not among the top variables of importance, implying that they were less positively associated with placement differences, even though these measures have shown significance in previous studies (e.g., Guo et al. 2013). As with the summary task, the model showed limited performance in drawing sharp distinctions between neighboring

placement levels, reinforcing the challenges of fine-grained automated classification using linguistic indices alone.

5.3 Comparison Between Summary and Argumentative Writing

In this section, the most important linguistic features found in the summary corpus and argumentative writing corpus are discussed. One index, word familiarity, was found to be an important feature in classifying students' summary and argumentative writing essays according to their placement levels. This would suggest that one of the strongest features that distinguishes lower-level writing (B and C) from higher-level writing (D and P) may be the extent to which writers frequently incorporate familiar content words. For both the summary and argumentative writing tasks, B- and C-level students tended to use more familiar words than D- and P-level students. For example, phrases such as "pick up social skill" and "spending time online," taken from a B-level essay, reflect the use of highly familiar, everyday vocabulary. In contrast, a P-level essay included expressions like "self-expression," "essential element of communication," and "portrayed in a negative view," indicating the use of more abstract and less familiar lexical items. Word familiarity, as was found in other studies (e.g., Crossley et al. 2011, Guo et al. 2013, Jung et al. 2019), can be a strong index for estimating L2 essay quality. In addition, the top-ranked predictors in both writing tasks were mostly related to lexical sophistication and sentence complexity. This too confirms what was found in studies (e.g., Crossley and McNamara 2012a, McNamara et al. 2015, Jung et al. 2019) using Coh-Metrix indices as a means to predict proficiency levels or essay quality. Moreover, the findings of the present study align with previous research that has found a significant relationship between writing placement decisions and lexical sophistication or syntactic complexity (e.g., Jin 2023, Kim 2020, Nguyen 2024, Vo 2019). While LDMTLD (as a measure of cohesion) appeared as one of the top five important variables classifying argumentative essays, none of the other indices related to cohesion appeared. This is not to say that cohesion as a writing construct was less important, but rather that its role in placement decisions may be more subtle or task-dependent. One possible explanation may be that writers across all levels could take advantage of the structures found in source texts and incorporate them into their own essays (Guo et al. 2013, Kyle 2020). In such cases, significant differences in cohesion may be more difficult to observe.

Although both tasks were integrated and source-based, the RF models revealed subtle genre-sensitive patterns in their top-ranked features. RF analysis on the summary corpus showed that word familiarity, age-of-acquisition scores, word meaningfulness, left embeddedness, and sentence syntax similarity were important variables for classification. Particularly, students placed at the lowest level (B) were found to use more high-frequency, meaningful words and simpler sentence structures. Students at the highest level (P), by contrast, used more advanced words displaying limited sense relations and meaningfulness, and their writing demonstrated more sentence complexity and variety. For instance, one sentence from the B-level read, "Besides that, social networking imporve human's attitude toward life," while the P-level corpus featured more abstract phrasing and deeper clause embedding, including, "Social media could be considered a double-edged sword that can has positive influence over people's lives or disastrous consequences depending on what it's use for." When it comes to writing an integrated source-based summary essay, then, the present study adds further support that it is important for students to clearly demonstrate their ability to use a range of syntactic (Casal and Lee 2019, Crossley and McNamara 2014) and vocabulary structures (Crossley et al. 2011).

RF analysis on the argumentative writing corpus showed word familiarity, lexical diversity, the average number of modifiers per noun phrase, word meaningfulness, and age-of-acquisition scores playing an important role in making placement level predictions. Along with word familiarity, higher values of word meaningfulness contributed more strongly to the classification of B- and C-level essays. In contrast, higher values of lexical

diversity and mean number of modifiers per noun phrase were more influential in classifying essays into D and P levels. For instance, the sentence, "They are light, and easy to use, and they are helpful in leisure seeking," from a C-level essay features simple coordination and repeated phrasing. On the other hand, a D-level sentence, "Students can file the steps to do a test, search in internet for articles related to that text, share data and information with each other," reveals multi-clause construction and denser informational packaging. In terms of model performance, slightly stronger recall and F1 scores were observed for the P-level group in the argumentative writing task, indicating that certain lexical and structural features in higher-level argumentative essays were more reliably captured by the model. In contrast, the summary writing model showed more balanced, but still limited, performance across levels, particularly in terms of adjacent accuracy. These results may reflect subtle genre-related differences. Argumentative writing at higher levels tended to feature greater lexical diversity and noun phrase complexity, likely due to the task's greater demand for logical and coherent claim development and dense information delivery. On the other hand, summary writing tended to show more gradual differences in lexical and syntactic features across placement levels, likely due to the task's focus on paraphrasing and organizing information from external sources. While these patterns do not suggest a sharp divide between the two genres, they do indicate that each writing task may bring out different distributions of linguistic features across placement levels. The current study extends previous studies by showing that a good display of lexical sophistication, syntactic complexity, and cohesion is important for performing well on integrated reading-writing tasks (Crossley et al. 2011, Guo et al. 2013). A key contribution of this study is that raters may be paying attention to slightly different aspects of lexical sophistication, syntactic complexity, and cohesion even across closely related academic tasks such as integrated summary and argumentative writing, and that good writers can manipulate such nuanced linguistic features effectively, a point that has been underexplored in previous placement test-based research. Unlike earlier studies relying on classical statistical methods such as Chi-square analysis (Vo 2019) or ANOVA (e.g., Nguyen 2024), this study has used a RF model to evaluate nonlinear relationships between linguistic features and placement decisions to obtain such nuanced observations. Finally, in terms of test validity, the findings cautiously add support to having two different kinds of source-based writing tasks on the English placement writing exam (Li 2015), though further research may be needed to fully confirm this argument.

6. Conclusion

This study separately identified the top five most important linguistic features that are predictive of students' summary essay writing qualities and argumentative writing qualities using Coh-Metrix and the RF method. Compared to previous studies that have yielded accuracy and precision at least above 60% using Coh-Metrix indices (e.g., Crossley and McNamara 2009, 2012a, Crossley et al. 2011), the RF models adopted in the present study did not perform as strongly. Although the adoption of RF was prompted by the poor performance of the QDA models, classification performance—including accuracy, precision, recall, and macro F1—remained modest overall despite the application of this more sophisticated algorithm. Even after applying SMOTE-Tomek resampling to address class imbalance, the models exhibited difficulty in distinguishing adjacent placement levels, suggesting that shared linguistic features across levels may have reduced the model's predictive power. The unbalanced class size, the relatively small sample size available for training, and the challenges of predicting placement levels of L2 writers who fall within a narrower proficiency range may explain some of the issues observed in the current study.

Despite these constraints, the RF models were able to detect broad patterns associated with placement levels,

particularly when comparing students at the lower placement levels (B and C) and those at the higher levels (D and P). In turn, this lends cautious support to the use of placement testing for grouping students into general placement levels. Although the models cannot identify precise individualized instructional needs, the findings of this study suggest that attention to lexical sophistication (Vögelin et al. 2019), syntactic complexity (Casal and Lee 2019), and cohesion (Crossley et al. 2016), may help inform broad instructional priorities, as these features appeared more consistently, though not exclusively, in higher placement level essays. What is more, genre-specific analyses revealed subtle differences in how linguistic features contributed to placement predictions. Argumentative essays placed greater emphasis on structured argumentation and noun phrase elaboration, while summary essays reflected more gradual variation in lexical and syntactic complexity across placement levels. These genre-sensitive patterns offer tentative support for including both task types in placement tests, as each may elicit different aspects of learner writing.

More broadly, these findings may help inform refining rubric descriptors as well as rater and teacher training. For instance, instead of relying on general statements such as "a wide range of grammar structures and vocabulary appropriately and accurately," rubrics might reference more specific indicators such as lexical diversity, noun phrase complexity, or sentence-level cohesion to guide holistic judgment. Raters/instructors may benefit from paying attention to specific features when rating or scaffolding instruction tailored to students at specific levels (e.g., are the words used by students generally more abstract/academic or concrete?).

A point worth mentioning is that Coh-Metrix is not representative of all the linguistic features that are taken into account when raters score timed essays. In the future, it would be useful to consider other important Coh-Metrix features that were not explored in the current study as well as rhetorical aspects of writing (Crossley 2020). Limited timing, essay length, grammar and spelling errors, the degree to which students used source texts verbatim, and topic variation may also have had an effect on human rating, and these should be taken into consideration in future studies as well. While human raters likely draw on rhetorical structure, coherence, and content relevance in ways not captured by automated indices, the classification patterns found in this study suggest that linguistic features can offer complementary insights into broader trends in placement decisions. Finally, a compilation of a larger, more balanced corpus would improve predictive performance and finer-grained analysis of placement-level distinctions. To conclude, this study has explored whether computational indices could help differentiate essays written at varying placement levels using a Random Forest classification approach. While the models' overall classification performance was modest, the findings highlight linguistic features, such as lexical sophistication and syntactic complexity, that likely contribute to rater judgments, suggesting that computational analyses can offer complementary perspectives on placement decisions in L2 writing.

References

Batista, G. E., R. C. Prati and M. C. Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 6(1), 20-29.

Biber, D. and B. Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9, 2-20.

Casal, J. E. and J. J. Lee. 2019. Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing* 44, 51-62.

Crossley, S. A. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11(3), 415-443.

Crossley, S. A. and D. S. McNamara. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing* 18, 119-135.

Crossley, S. A. and D. S. McNamara. 2012a. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S.A. Crossley and S. Jarvis, eds., *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*, 106-126. Multilingual Matters.

Crossley, S. A. and D. S. McNamara. 2012b. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2), 115-135.

Crossley, S. A. and D. S. McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26, 66-79.

Crossley, S. A. and D. S. McNamara. 2016. Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research* 7(3), 351-370.

Crossley, S. A., K. Kyle and D. S. McNamara. 2016. The development and use of cohesive devices in L2 writing and their relations to essay quality. *Journal of Second Language Writing* 32, 1-16.

Crossley, S. A., T. Salsbury and D. S. McNamara. 2011. Predicting the proficiency level of language learners using lexical indices. *Language Testing* 29(2), 243-263.

Danis, N. 2019. *Variation of Linguistic Markers of Stance in ESL students' Summary and Argumentative Essays*. Master's thesis, Iowa State University.

Goh, T-T., H. Sun and B. Yang. 2020. Microfeatures influencing writing quality: The case of Chinese students' SAT essays. *Computer Assisted Language Learning* 33, 455-481.

Guo, L., S. A. Crossley and D. S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18, 218-238.

ISU EPT Corpus of Learner Writing (Release 2.2). 2017. Corpus compiled by the Applied Linguistics and Technology program and Bethany Gray at Iowa State University.

Jin, H. 2023. Lexical frames and errors in the use of English definite article in L2 academic writing: A case of English placement test. *Korean Journal of English Language and Linguistics* 23, 324-341.

Jung, Y., S. A. Crossley and D. S. McNamara. 2019. Predicting second language writing proficiency in learner texts using computational tools. *The Journal of Asia TEFL* 16(1), 37-52.

Kim, H. 2020. Nominal modifiers in argumentative essays as discriminators for writing course placement decisions. *English Teaching* 75(3), 3-24.

Kyle, K. 2020. The relationship between features of source text use and integrated writing quality. *Assessing Writing* 45, 100567.

Kyle, K. and S. A. Crossley. 2016. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing* 34, 12-24.

Li, Z. 2015. *An Argument-Based Validation Study of the English Placement Test (EPT): Focusing on The Inferences of Extrapolation and Ramification*. Doctoral dissertation, Iowa State University.

Ma, H., J. Wang and L. He. 2024. Linguistic features distinguishing students' writing ability aligned with CEFR levels. *Applied Linguistics* 45, 637-657.

McNamara, D. S., S. A. Crossley, R. D. Roscoe, L. K. Allen and J. Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23, 35-59.

McNamara, D. S., A. C. Graesser, P. M. McCarthy and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge.

Mizumoto, A. 2023. Calculating the relative importance of multiple regression predictor variables using

dominance analysis and random forests. *Language Learning* 73(1), 161-196.
Nguyen, P. 2024. Noun phrase complexity in English integrated writing placement test responses. *Journal of English for Academic Purposes* 72, 101452.
R Core Team. 2025. R: A language and environment for statistical computing (version 4.3.1) [Computer software]. Available online at https://www.R-project.org
Vo, S. 2019. Use of lexical features in non-native academic writing. *Journal of Second Language Writing* 44, 1-12.
Vögelin, C., T. Jansen, S. D. Keller, N. Machts and J. Möller. 2019. The influence of lexical features on teacher judgments of ESL argumentative essays. *Assessing Writing* 39, 50-63.
Zagata, E., D. Kearns, A. J. Truckenmiller and Z. Zhao. 2023. Using the features of written compositions to understand reading comprehension. *Reading Research Quarterly* 58(4), 624-654.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Appendix A. Partial Dependence Plots of the Summary Corpus

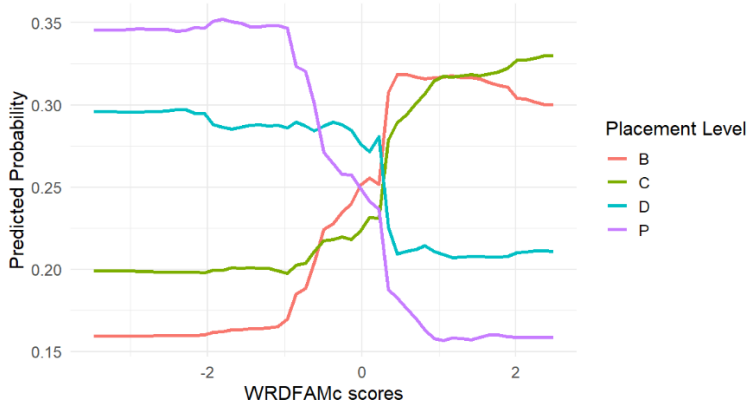

Figure A1. Partial Dependence of Placement Predictions on WRDFAMc (Summary Corpus)

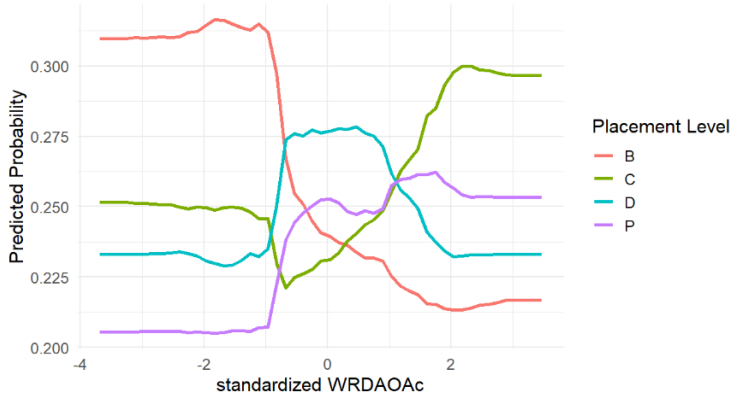

Figure A2. Partial Dependence of Placement Predictions on WRDAOAc (Summary Corpus)

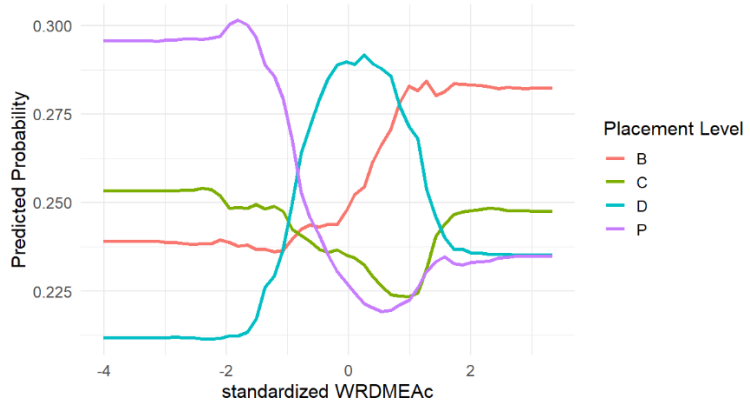

Figure A3. Partial Dependence of Placement Predictions on WRDMEAc (Summary Corpus)

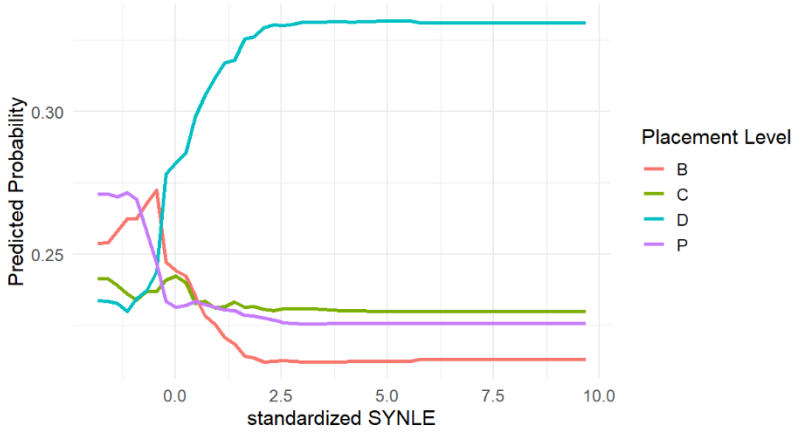

Figure A4. Partial Dependence of Placement Predictions on SYNLE (Summary Corpus)

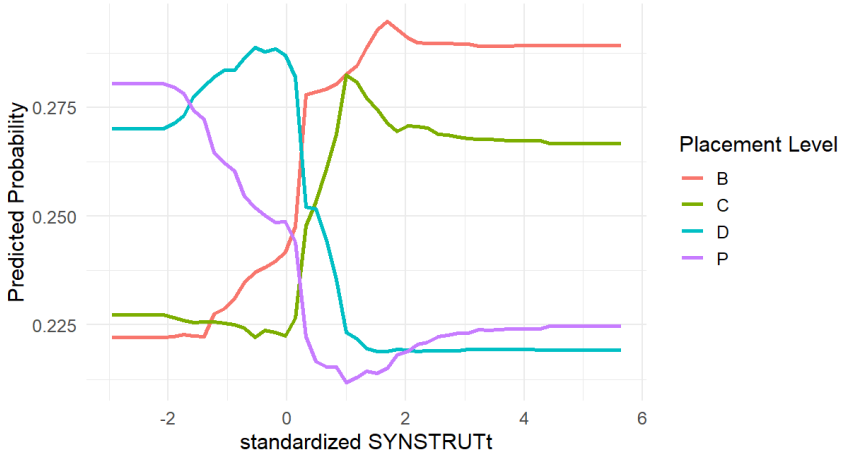

Figure A5. Partial Dependence of Placement Predictions on SYNSTRUTt (Summary Corpus)

Appendix B. Partial Dependence Plots of the Argumentative Writing Corpus

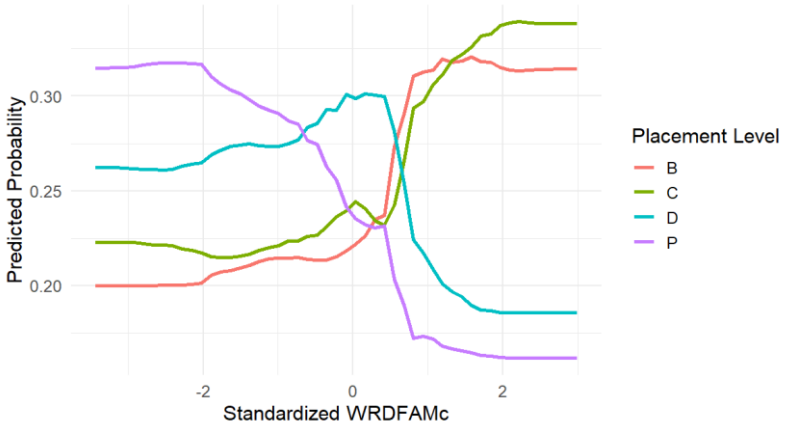

Figure B1. Partial Dependence of Placement Predictions on WRDFAMc (Argumentative Corpus)

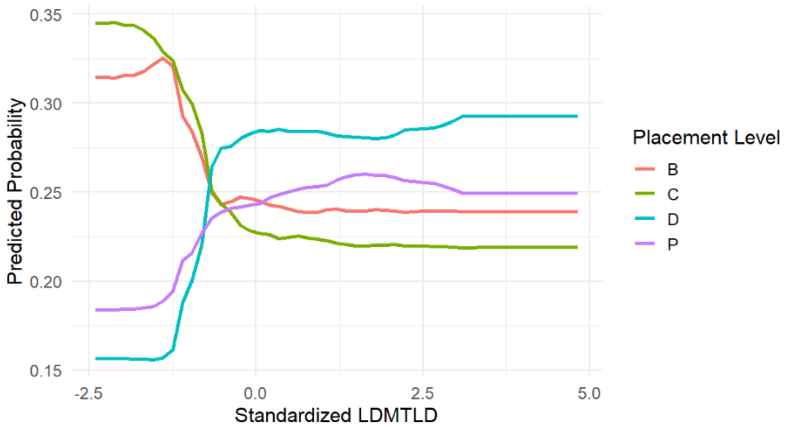

Figure B2. Partial Dependence of Placement Predictions on LDMTLD (Argumentative Corpus)

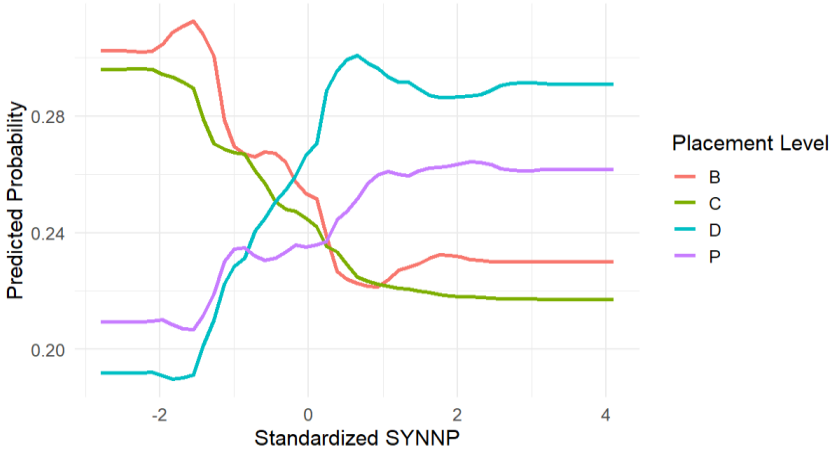

Figure B3. Partial Dependence of Placement Predictions on SYNNP (Argumentative Corpus)

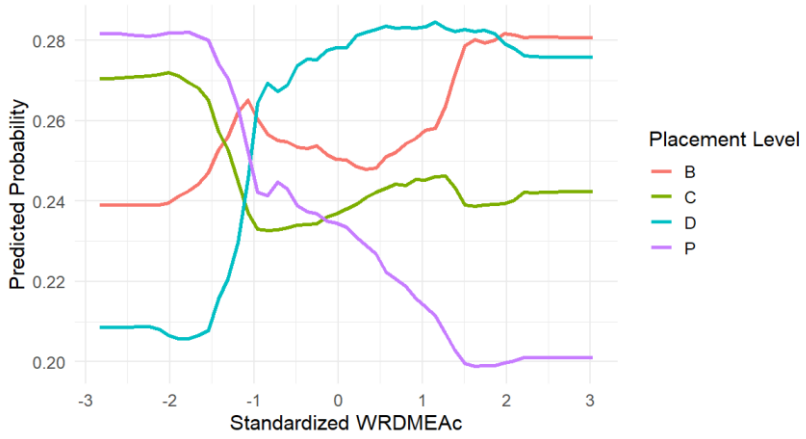

Figure B4. Partial Dependence of Placement Predictions on WRDMEAc (Argumentative Corpus)

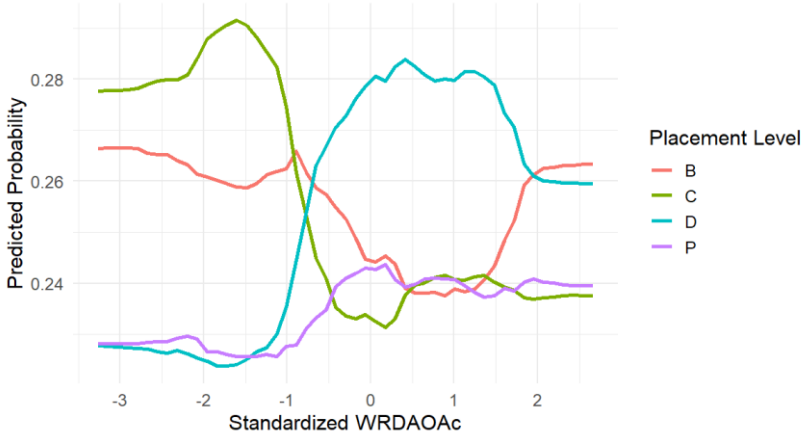

Figure B5. Partial Dependence of Placement Predictions on WRDAOAc (Argumentative Corpus)