Enhancing Writing Proficiency Classification in Developmental Education: the Quest for Accuracy

Miguel Da Corte[1,2], **Jorge Baptista**[1,2]
[1]University of Algarve, [2]INESC-ID Lisboa
[1]Faro, [2]Lisbon (Portugal)
miguel.dacorte@tulsacc.edu, jbaptis@ualg.pt

Abstract

Developmental Education (DevEd) courses align students' college-readiness skills with higher education literacy demands. These courses often use automated assessment tools like ACCUPLACER for student placement. Existing literature raises concerns about these exams' accuracy and placement precision due to their narrow representation of the writing process. These concerns warrant further attention within the domain of automatic placement systems, particularly in the establishment of a reference corpus of annotated essays for these systems' machine/deep learning. This study aims at an enhanced annotation procedure to assess college students' writing patterns more accurately. It examines the efficacy of machine-learning-based DevEd placement, contrasting ACCUPLACER's classification of 100 college-intending students' essays into two levels (Level 1 and 2) against that of 6 human raters. The classification task encompassed the assessment of the 6 textual criteria currently used by ACCUPLACER: mechanical conventions, sentence variety & style, idea development & support, organization & structure, purpose & focus, and critical thinking. Results revealed low inter-rater agreement, both on the individual criteria and the overall classification, suggesting human assessment of writing proficiency can be inconsistent in this context. To achieve a more accurate determination of writing proficiency and improve DevEd placement, more robust classification methods are thus required.

Keywords: developmental education, machine learning, natural language processing, corpus annotation methods

1. Introduction and Objectives

This paper addresses students' writing skills (with English as their L1) in view of their placement in a two-level, DevEd course model. Within this model, students remediate linguistic deficiencies until they reach a level of written language proficiency sufficient for them to aptly participate in an academic program. Placement in DevEd is often based on the automated assessment and scoring of linguistic features extracted from standardized written assignments, administered as part of an entrance exam, and using automatic systems, such as ACCUPLACER.[1]

According to Hassel and Giordano (2015); Nazzal et al. (2020), standardized exams, like the one mentioned, demonstrate some limitations in the classification precision and portray a narrow conceptualization of the writing process (Hughes and Li, 2019). It is estimated that these standardized exams misplace about 30% to 50% of students (Hassel and Giordano, 2015) who complete the exams prior to beginning their college career. Furthermore, "test scores poorly correlate with students' ultimate success in college, particularly at campuses that rely on standardized placement tests to sort students into appropriate coursework." (Hassel and Giordano, 2015, p.58).

At Tulsa Community College (TCC)[2], the higher education institution where this study took place, ACCUPLACER is the automated system used to determine a student's writing proficiency level and, consequently, their placement in DevEd or college-level writing courses. Although this assessment is used nationwide, it is pertinent to note that guidelines to participate in DevEd courses in the United States of America vary from one state to the other, and, in some instances, from one educational institution to another. Thus, this variation justifies the need to continue investigating ways to improve how these systems perform to effectively provide access and placement to educational opportunities for students who may be underprepared, particularly at the community college level.

In view of these limitations and their impact on students' course placement, this study aims at (i) identifying the linguistic features that are more predictive of placement in DevEd, (ii) assessing and selecting relevant features that could contribute to achieving a language proficiency equivalent to that of native English speakers, and (iii) supporting a more systematic placement of students by enhancing automatic classification systems.

[1] https://www.accuplacer.org/ (Last access: March 21, 2024; all URLs in this paper were checked on this date.)

[2] https://www.tulsacc.edu/

LREC-COLING 2024, pages 6134–6143
20-25 May, 2024.

2. Related Work

Higher education aims to provide a path to literacy through DevED, particularly in community colleges in the United States (Mazzariello et al., 2018; Cormier and Bickerstaff, 2019; Bickerstaff et al., 2021). In English L1, specifically, DevEd courses are designed to strengthen students' competencies in reading and writing. However, the placement process in DevEd courses has been an object of debate, primarily due to concerns about the validity and accuracy of test results (Perin et al., 2015; Barnett et al., 2020), which has also raised issues of ethics, fairness, and equity (Holmes et al., 2021; Porayska-Pomsta and Rajendran, 2019; Denison-Furness et al., 2022).

For students to persist and progress academically, proficiency in reading and writing is essential. Therefore, it is crucial to identify and understand the linguistic features of developing writers, at a more granular level, with the ultimate goal of proper placement into (and support throughout) DevEd or college-level courses. Furthermore, understanding the writing patterns of community college students illuminates the language-related issues that interfere with effective communication in higher education. Several studies (Ramesh et al., 2011; Pal and Pal, 2013; Zhang and Aslan, 2021) have relied on the use of data mining methodologies to correlate students' placement in academic courses and other college-related activities with their academic performance and college success.

Zhu et al. (2020) used NLP tools and automated scoring systems to assess students' written argument production and provide feedback on how to improve writing skills when supporting topics, even in complex areas such as sciences. The correlation of feedback to performance gains in writing skills was also explored. Results suggested that "appropriate scaffolding and feedback for students can be further developed depending on the nature of the tasks and the performance of students (Zhu et al., 2020, p.12).

Analyzing the lexical and syntactical patterns that students exhibit early in their developmental education trajectory is helpful in guiding teaching practices that support proficiency needs, particularly as these linguistic features can be best addressed and taught by reading and writing for a specific and meaningful social goal (Dowell and Kovanovic, 2022). Many of the students who participate in DevEd courses are highly concerned with vocabulary and how to produce a narrative that is structured, congruent, and functional. More precisely, they are concerned with the lexical richness that captures their active vocabulary size and has the potential to impact their academic performance (Hilte et al., 2020).

Consequently, identifying more descriptive linguistic features and integrating them in the training of systems like ACCUPLACER could optimize classification by skill level, setting the stage for the current study.

3. Methodology

3.1. Corpus

A sample of 103 text units (essays), written in English, was randomly selected from texts produced by a population of 290 college-intending students during the 2021-2022 academic year. These essays were written in an unrestricted time frame and without the availability of editing resources at TCC's supervised testing facility. The samples were extracted from the institution's standardized entrance exam database in plain text format, strictly adhering to the protocols for human subject protection. The primary metadata denoted students' DevEd placement level (Level 1 or Level 2) as determined by ACCUPLACER. Other metadata, including demographics (gender, race, among others), was ignored at this stage. Table 1 shows the corpus statistics.

Corpus	Total
Tokens	27,916
Average tokens per text	279
Maximum number of tokens in a text	422
Minimum number of tokens in a text	95

Table 1: Corpus statistics.

Text unit samples were evenly distributed between the two placement levels, but varied in length from 95 to 422 words. To address this imbalance in the corpus, units were divided into 100-word segments, and a random resampling was performed to ensure balance across levels (50 for Level 1 and 50 for Level 2). For the random resampling, a random numbers table was utilized to ensure a fair and unbiased selection of split samples from the corpus.

The number of the resampled text units was the same: 103. Of these sample units, 3 were used for the training of the annotators, while 100 were used for the core annotation task. This subset serves as the foundational corpus for the study of the language skills of community college students, particularly native English speakers. It will also aid in creating an annotation scheme to pinpoint significant linguistic characteristics of this population.

3.2. Classification Task

Students' placement is construed here as a classification task to assess the adequateness of machine-learning-based DevEd placement. To achieve this, the Accuplacer classification of the 100 text units is compared with the classification produced by a group of trained raters, who manually evaluated the same sample texts.

The classification task involved two main steps, in which raters (i) assessed each essay based on the 6 textual criteria (presented in Section 3.3) and (ii) produced an overall classification by skill level corresponding to DevEd Level 1, 2, or College level (demonstrating no need for DevEd). After conducting experiments with logistic regression using ordinal classes, results did not significantly differ from those obtained with nominal classification. As a result, nominal classification was adopted.

Prior to deploying the classification task, two linguists, with extensive experience in signaling and categorizing linguistic features in written corpora, carried out a pilot classification task and annotated a few randomly selected texts from the same database and period, based on the features currently used by Accuplacer.

This pilot allowed the opportunity to investigate any potential discrepancies regarding the difficulty of the task at hand. After this pilot, a consensus was reached on the interpretation and use of the 6 textual criteria (discursive patterns) reportedly adopted by said system (The College Board, 2022), allowing the development of the annotation guidelines presented in the next section.

3.3. Annotation Guidelines

For the purposes of this study, *discursive patterns* are defined as the patterns exhibited in the production of multiple utterances or extended texts to discuss a topic, reformulate it, support an opinion, and hypothesize, among other higher-order thinking tasks.

Based on this definition, the 6 features on which this annotation task focuses are: (i) *Mechanical Conventions*; (ii) *Sentence Variety & Style*; (iii) *Idea Development & Support*; (iv) *Organization & Structure*; (v) *Purpose & Focus*; and (vi) *Critical Thinking*. Within each textual criterion, a 4-point Likert scale was adopted: 0 - *deficient*; 1 - *below average*; 2 - *above average*; 3 - *outstanding*.

The definitions provided by the literature on Accuplacer are scarce. However, as part of this experiment, the two linguists illustrated these definitions with real examples – something that the (The College Board, 2022) does not provide – and created a document with guidelines[3] for training purposes (Da Corte and Baptista, 2024b). All 6 raters received a copy of this document, with instructions on how to use it, to ensure that the construct of each feature was clearly understood and consistently applied throughout the task.

An excerpt of how the features were presented to the raters is provided below:

> Definition based on accuplacer's training manual (The College Board, 2022, p.27)[4]:
> *Mechanical Conventions* refer to the extent to which the text expresses ideas using standard English.

The scope of the feature is then provided (guidelines authors' wording):

> As you rate the text within this category, consider the following elements: spelling, grammar, and punctuation.

With the 4-point Likert adopted for each category[5].

> In terms of *Mechanical Conventions*, the text is (choose as appropriate): 0 - deficient; 1 - below average; 2 - above average; 3 - outstanding.

Selecting, for example, a score of 1 (below average) for Mechanical Conventions, indicates that the text read:

> (i) Had several typos. Scale to be used: 0 - deficient: 15 or more typos; 1 - below average: 8-14 typos; 2 - above average: 1 to 7 typos; 3 - outstanding: no typos.
>
> (ii) Evidenced run-on sentences throughout, e.g., "*I agree with that statement because I know about changing myself, I went from a depressed misrable woman who weighted alomst 500lbs to a driven, happy independent woman who weighs 300lbs and is still loosing weight[...].*"
>
> (iii) Used (or not used at all) punctuation signs incorrectly, e.g., "*My father was never in the picture much and [,] when he was [,] he constantly told me how I was gonna end up just like my mother[.] She who was a young mother depending on others to help take care of her children.*"
>
> (iv) Included the use of contractions, e.g., *didn't; I'm*; or slang, e.g., *gonna; wanna; gotta*, appear on the text. These linguistic features are not used in common academic writing.

For the classification by level, the following definitions were adopted:

[3] https://gitlab.hlt.inesc-id.pt/u000803/deved/

[4] The experiments were conducted prior to the current (2022) accuplacer Program Manual's publication (The College Board, 2022), which further elaborated on the 2018 edition (The College Board, 2018). While the 2022 edition expanded on score definitions and descriptors and provided detailed assessment statements, these changes have no impact on the experiments conducted or on the reported results.

[5] The 8-point Likert scale of The College Board (2022, pp.24 ff.) was too complex for the task at hand. Therefore, the 4-point scale, here developed and adopted, consists of descriptors aligned with DevEd terminology.

Level 1 DevEd: if the text indicated that development was needed in the overall use of the English language: grammar, spelling, punctuation, and sentence and paragraph structure.

Level 2 DevEd: if the text suggests that support is needed in specific areas versus the overall use of the English language, e.g., sentence structure, punctuation, editing, and revising.

Level 3 College-level: indicating that the text is written accurately and showcases the use of proper English at the college/academic level (the text successfully communicates and supports specific ideas or points of view).

3.4. Annotators and Training

A call for volunteers to participate in the annotation task was disseminated at TCC and local partner community agencies. A total of 6 participants responded to the call, and all were selected based on their background, skills, and experience.

Demographic information from the raters was requested and can be summarized as follows: (i) Gender: 33% female, 67% male; (ii) Language Background: 83% native English speakers, 17% English as a second language (with at least a near level of English proficiency); (iii) Education: 33% hold at least a Bachelor's degree, 67% have a Master's degree; (iv) Self-rated English skills: 83% advanced, 17% superior; (v) Employment: 83% work in higher education, 17% work in the private sector.

Raters were trained on how to apply the set of annotation guidelines to the writing samples collected. The training, provided by one of the guidelines' authors, covered the expectations, ethical considerations, steps of the annotation task, and the annotation guidelines. Two practice rounds of annotations were completed as part of the training, one *guided* and one *independent*. After the *independent* round was completed, a debrief session was scheduled where a sample of the annotations and disagreements among them were discussed.

A timeline was established for the annotation task to be completed within 15 days, with a midway check-in scheduled for day 7. Both the training session and the annotation of the 100 sample units were uncompensated. The annotation began immediately after the debrief.

4. Annotation Task Assessment

All 100 essays were assessed within the allotted timeline. On average, raters self-reported spending 13 minutes per essay.

Text samples were randomly assigned to raters, with approximately 49 essays assessed by at least two raters each. One essay was discarded due to technical issues. Having essays assessed by at least two raters resulted in a combination of 5 pairs of raters. These 49 essays are the focus of the annotation assessment here explained. The strategy of limiting the number of essays per annotator was adopted to manage the significant task workload effectively. By ensuring each text unit was annotated by at least two different raters, the quality of the annotations was maintained without overburdening the raters. Intraclass Correlation Coefficients (ICC) (Koo and Li, 2016) and Krippendorff's Alpha (K-alpha) Interrater Reliability Coefficient (IRC) were calculated using the ReCal-OIR tool (Freelon, 2013)[6], to provide insights into the agreement between raters within each pair.

Regarding the classification by skill level of these 49 essays, 27 (55%) received the same classification level from two raters; 22 (or 45%) received a different classification level. Within these 22 essays, 12 were placed at a level higher than the one suggested by ACCUPLACER; 9 were placed at a level below, while only 1 essay was classified two levels apart from the class indicated by ACCUPLACER (Level 1 versus College level). These preliminary results call for a closer inspection of the interrater reliability, which is the purpose of Section 4.1. The dataset with the respective scores (integers) for all 49 essays can be found on Da Corte and Baptista (2024a).[7]

4.1. Intraclass Correlation Coefficients

To assess the reliability of the 6 linguistic features, ICC were calculated for the 49 essays mentioned in Section 4. A two-factor ANOVA, at 95% confidence level without replication, was used to calculate each pair's ICC. The results are summarized in Table 2, with the best ICC scores per feature and skill level in italics and the best scores per performing team in bold.

Low ICC scores were expected, especially for complex (not easy to define or capture) linguistic features like *Sentence Variety* and *Critical Thinking*. Conversely, features such as *Mechanical Conventions* and *Organization & Structure* were expected to achieve higher agreement between the annotators since they are more objective in nature. Nevertheless, lower ICC scores could have derived from the raters' experiences and background, e.g., views on DevEd, the complexity of the task as a whole, among others.

For the interpretation of ICC scores, the coefficient thresholds and interpretations set forth by Koo and Li (2016) were followed. The authors claim that potential low ICC scores are attributed to the "lack of variability among the sampled subjects, the small

[6] http://dfreelon.org/recal/recal-oir.php

[7] https://gitlab.hlt.inesc-id.pt/u000803/deved/

Linguistic Features	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
Mechanical Conventions	**0.533**	**0.394**	*0.778*	0.111	0.360
Sentence Variety	0.364	0.111	0.111	0.423	***0.448***
Idea Development & Support	0.363	0.273	*0.814*	0.333	-0.294
Organization & Structure	0.203	-0.333	***0.849***	0.385	0.391
Purpose & Focus	***0.533***	-0.241	0.529	-0.412	0.220
Critical Thinking	0.276	*	0.707	***0.857***	*
Skill Level	0.176	0.111	***0.789***	0.077	0.030

Table 2: ICC for all 5 pairs of raters based on 6 linguistic features and classification by skill level.

number of subjects, and the small number of raters being tested" (Koo and Li, 2016, p.158). With this caveat in mind, this study employed more than 30 heterogeneous samples (n= 49) and involved a minimum of 3 raters (6 raters were involved) were followed, which was suggested by the authors.

Results indicate mostly poor reliability, with ICC scores generally below 0.5. When analyzing the results per pair and per feature, Pair 1's ICC scores were moderately reliable for the *Mechanical Conventions* and *Purpose & Focus* for a set 9 texts that this pair had in common.

Pair 2's had 10 essays in common. ICC scores in each of the features and skill levels were below 0.50. The *Mechanical Conventions*' ICC score was the highest with 0.394, but still indicated poor reliability. An extremely low value (-3.99e-16) was observed in this pair's evaluation of *Critical Thinking*, prompting additional calculations and comparisons using Krippendorff's alpha reliability coefficient (See Section 4.2). This result was ignored.

The performance exhibited by Pair 3, with 10 essays, presented more consistency both in assigning the rates to the six linguistic features and the overall skill level. With this pair, most features were within the good reliability range (0.75 - 0.90), except *Critical Thinking* and *Purpose and Focus*, which yielded moderate reliability (0.707 and 0.529, respectively), leaving *Sentence Variety* with a poor reliability ICC score of 0.111. Raters within this pair operated under the same context and circumstances as the other pairs. During the experiment, they served as academic advisors at TCC and were familiar with the Accuplacer system and its scoring. They also self-reported an understanding of the content taught in DevEd courses and the complexities of student placement processes.

It is important to note that besides the enhancements made to *Mechanical Conventions*, as mentioned in Section 3.3, the definition of *Organization & Structure* was also refined (by the guidelines authors), now including a statement referencing the main components of an essay: introduction, body paragraphs with supporting evidence, and conclusion. The precision of this enhancement could have played a role in the reliability of the results reported by this pair for this specific feature.

The last two pairs, Pairs 4 and 5, each had 10 essays in common and exhibited similar, low ICC scores in *Sentence Variety & Style* and *Organization & Structure* and in the Skill Level category. Scores, again, indicate poor reliability, with a score below 0.50. Despite these results, pair 4 had the best ICC score for *Critical Thinking* (0.857), suggesting good reliability. Like Pair 2, above, an outlier value (6.66e-16) was obtained by Pair 5 on the feature *Critical Thinking*. This result was also ignored.

Taking into consideration the linguistic features adopted and the corresponding descriptions, based on Accuplacer manual guidelines, and the ranking of each feature using a 4-point (ordinal) Likert scale, overall, results showed poor reliability in the rating process. Poor correlation scores were also observed, as shown in Table 3, when performing additional ICC calculations, comparing each rater (in each pair) with the classification assigned by Accuplacer. This poor correlation could be attributed to the narrow conceptualization of the writing process automated systems portray, according to Hassel and Giordano (2015); Nazzal et al. (2020).

Because of these limitations and the consequences this has on students' course placement, producing an extended annotation procedure that is more representative of the linguistic patterns of college students is paramount.

Pair	Rater 1	Rater 2
Pair 1	0.160	0.102
Pair 2	-0.111	**0.368**
Pair 3	**0.333**	0.052
Pair 4	0.158	0.333
Pair 5	-0.167	0.030

Table 3: ICC for all five (5) pairs versus ACCUPLACER in terms of skill classification level.

A final ICC calculation was performed to determine how human classification (all individual ratings) correlated with that of ACCUPLACER. A score of 0.119 was obtained, evidencing a poor correlation between humans and this automated system in the assessment of all 6 linguistic features.

4.2. Additional Reliability Calculations

In view of results obtained in Section 4.1, Interrater Reliability Coefficient (IRC) calculations were performed next, and results are presented in Table 4. The ReCal OIR tool was, again, used since it provides the calculation of Krippendorff's Alpha (K-alpha) for ordinal data for at least two raters. The best reliability scores per feature and skill level are italicized in Table 4, with the best scores per performing team in bold font.

For the interpretation of K-alpha scores, the thresholds and interpretation guidelines set forth by Cohen (1988) were followed. Results with K-alpha scores are very similar to the ICC scores obtained. However, with the K-alpha method, Pair 1 achieved moderate agreement (0.528) with *Mechanical Conventions*. An almost perfect agreement was reached by Pair 3 with *Organization & Structure* (0.842), and by Pair 4 with *Critical Thinking* (0.933). Pair 3 is still the best-performing, more consistent, team in this classification task. No outliers were observed with these new calculations. With the K-alpha method, more interpretable scores were obtained for *Critical Thinking* with Pair 2 and Pair 5, 0.006 and 0.161, respectively, still pointing to a none-to-slight agreement.

A final K-alpha calculation was performed comparing all individual ratings to determine how human classification equates to that of ACCUPLACER. A K-alpha score k=0.139 was obtained, evidencing a discrepancy between humans and this automated system when it comes to accurately measuring language proficiency and determining the need for DevEd (based on the classification level assigned). This score is not very different from the one obtained with ICC (0.199).

Pearson's correlation calculations between the ICC and the K-alpha scores were performed. An overall Pearson coefficient of r=0.862 and an average (pairwise) coefficient of 0.728 were obtained. Pairwise, Pair 3 achieved a r=0.991 (almost perfect correlation), with the lowest value of r=0.436 found for Pair 5. Even if some detailed analysis would be in order to account for the difference in the results, these high correlation values mean that, in general, the inter-rater agreement is low, confirming the poor reliability of the rating task completed.

In the next section, the essay ratings provided by the annotators were used to model the classification procedure, attempting to understand the relevance of each feature in the process and their relation with the overall classification in DevEd courses. The primary value of Section 5 lies in demonstrating how the classification task performed by ACCUPLACER compares to that of human annotators following the same guidelines this system uses.

5. Machine-learning Modeling

The data for the machine-learning experiments consisted of the ratings provided by 2 independent raters for each essay. As indicated in Section 4, 49 texts received 2 independent ratings (totaling 98 instances) for the 6 features and an overall skill classification of the essay by level: DevEd Level 1 and Level 2, and College level (not requiring DevEd). In this subset, there was no essay assigned to College level (in the total corpus, only 1 essay was classified into College level).

For this experiment, only the 27 essays (55%) on which both raters (per pair) agreed on the overall classification level were selected. Pearson scores were calculated to compare the correlation between the skill level classification assigned by humans and ACCUPLACER, for this particular subset. The main purpose of this experiment is to determine which are the most relevant features, to differentiate among Level 1 DevEd, Level 2 DevEd, and College Level, based on the classification produced by participating raters.

The data mining tool ORANGE (Demšar et al., 2013)[8] was selected for the analysis and modeling of the students' essays assessment. Figure 1 shows the workflow adopted. The RANK widget, with the built-in *Information Gain* and *Chi-square* (χ^2) scoring methods, was used to score the classification.

The workflow can be described as follows: the data is imported into Orange using the File widget – one line per essay (27*2=54), 6 columns for the features, plus a column with the overall skills level as the target variable. Data is then passed into the DATA SAMPLER widget that was configured to partition it for a 3-fold cross-validation, considering the small size of the sample, leaving 2/3 (36 instances) for training and 1/3 (18 instances) for testing purposes. Due to the dataset's configuration, stratified cross-validation is not feasible.

The TEST & SCORE widget was then used to determine the best-performing model. Four, different types of commonly used learning algorithms were selected: [Decision] Tree (DT), Random Forest (RF), Naive Bayes (NB) and Neural Network (NN). For the hyperparameters of the NN and RF machine-learning algorithms, the default values provided by the ORANGE text mining platform were employed. The results from this first part of the experiment are shown in Table 5.

The overall results of the learning step are quite high for all models. The NB learner ranked at the

[8] https://orangedatamining.com/

Linguistic Features	**Pair 1**	**Pair 2**	**Pair 3**	**Pair 4**	**Pair 5**
Mechanical Conventions	**0.528**	0.127	*0.715*	-0.108	**0.305**
Sentence Variety	*0.365*	-0.147	0.041	0.240	0.265
Idea Development & Support	0.141	**0.144**	*0.834*	0.348	0.015
Organization & Structure	0.206	0.015	***0.842***	0.128	-0.002
Purpose & Focus	*0.237*	-0.231	0.409	-0.343	0.106
Critical Thinking	-0.041	0.006	0.689	***0.933***	0.161
Skill Level	-0.104	0.240	*0.791*	0.054	0.184

Table 4: K-ALPHA IRC FOR ORDINAL DATA.

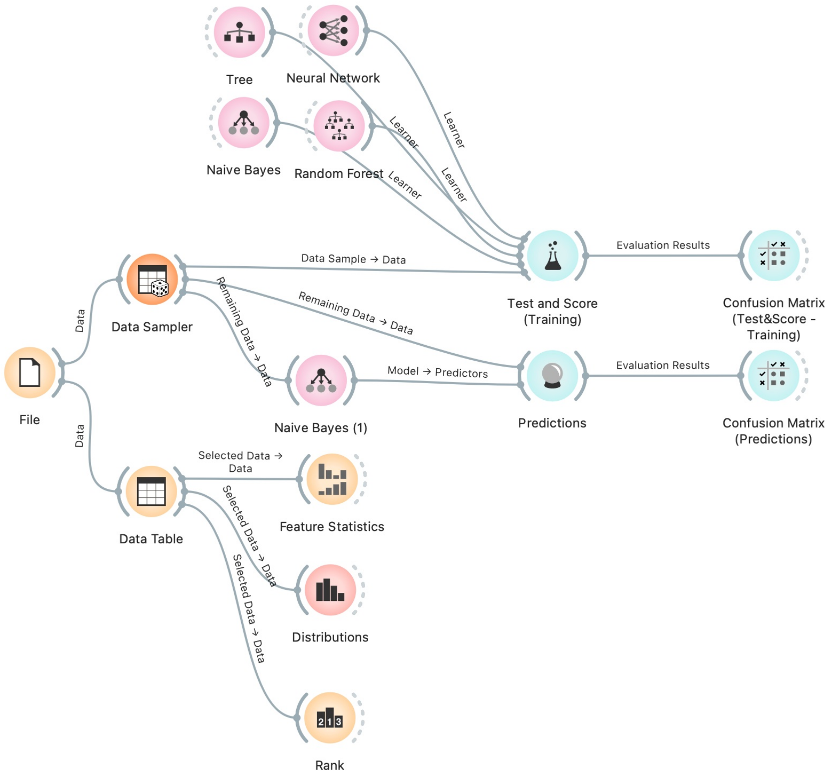


Figure 1: ORANGE Workflow Configuration for Model Training and Testing.

Model	**AUC**	**CA**	**F1**	**P**	**R**
NB	0.864	**0.917**	0.916	0.917	0.917
NN	0.851	0.833	0.835	0.842	0.833
RF	0.792	0.833	0.831	0.833	0.833
DT	0.774	0.833	0.831	0.833	0.833

Table 5: CLASSIFICATION ACCURACY PER DEVED LEVEL W/ 6 LINGUISTICS FEATURES (TRAINING).

top with a classification accuracy (CA) of 0.917. The other three learners, NN, RF, and DT, all ranked second, *ex aequo*, with a CA=0.833. However, their performance is not identical, as the models have differing Area under the ROC Curve (AUC) scores. For NB, AUC=0.864; 0.851 for NN, 0.792 for RF, and 0.772 for DT. In this case, the NN's AUC (0.851), is higher than that of RF and DT but not as good as NB's (0.864).

The system then ran the *Predictions* widget on the remaining 1/3 of the (unseen) data. With this smaller sample, results show a slight increase in all models' overall performance, except NB. This model ranked highest in the training phase. Still, a different ranking of the models was produced. While in the training step, the ranking order was:

NB > NN > RF > DT,

in this testing step, the order is:

NN > RF > DT > NB.

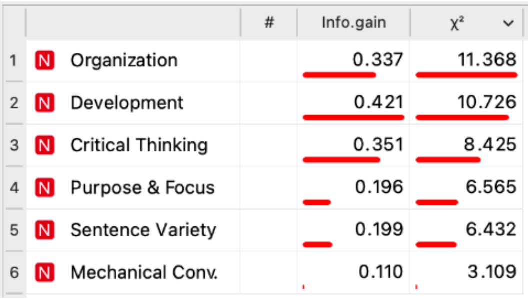
This almost complete reversal of the models' ranking order suggests that the sampling procedure may not be producing entirely consistent results. These issues are to be addressed in future work.

Results from this second part of the experiment are shown in Table 6.

Model	AUC	CA	F1	P	R
NN	0.993	**0.944**	0.943	0.949	0.944
RF	0.938	**0.944**	0.943	0.949	0.944
DT	0.958	0.889	0.882	0.905	0.889
NB	0.938	0.778	0.784	0.808	0.778

Table 6: Classification Accuracy per DevEd Level, w/ 6 linguistics features (testing).

Following these results, the Rank widget was then used to score the linguistic features according to their correlation with the target variable (Skill Level), based on applicable internal scores. In this case, the following scoring methods were used: (i) *Information Gain*, which indicates the expected amount of information (reduction of entropy); and (ii) *Chi-square* (χ^2), which shows the dependence between the feature and the class. Results are shown in Figure 2.



		#	Info.gain	χ²
1	N Organization		0.337	11.368
2	N Development		0.421	10.726
3	N Critical Thinking		0.351	8.425
4	N Purpose & Focus		0.196	6.565
5	N Sentence Variety		0.199	6.432
6	N Mechanical Conv.		0.110	3.109

Figure 2: Orange Feature Ranking (*Chi-square*).

The best ranking features in correlation with the target variable, using the *Information Gain* score, are *Idea Development & Support*, *Critical Thinking*, and *Organization & Structure*. The *Sentence Variety & Style* and *Purpose & Focus* features have smaller *Information Gain* values, thus contributing to the classification much less than the 3 features above. Within these values, *Mechanical Conventions* is the least informative feature. The *Chi-square* (χ^2) corroborates the previous two scores (with a different ranking of the 3 topmost features).

As shown in Table 7, by using the data sample plus the 3 topmost ranked features for training, the Test & Score widget results indicate that the performance of the same models deteriorates slightly, in terms of CA, for NB and DT, while increasing minimally for NN and RF. Though, the area under the curve (AUC) improves, when compared with the values of Table 5, for all the models except DT.

Like in the previous experiment, the remaining unseen data (1/3) was tested using the Predictions widget. Results are shown in Table 8.

Model	AUC	CA	F1	P	R
RF	0.935	**0.889**	0.889	0.889	0.889
NN	0.912	0.861	0.860	0.860	0.861
NB	0.919	0.833	0.831	0.833	0.833
DT	0.638	0.694	0.664	0.699	0.694

Table 7: Classification Accuracy per DevEd Level, 3 Topmost Ranked Features (training).

Model	AUC	CA	F1	P	R
RF	0.964	**0.889**	0.884	0.906	0.889
NB	0.964	**0.889**	0.889	0.889	0.889
NN	0.958	**0.889**	0.889	0.889	0.889
DT	0.911	0.861	0.853	0.887	0.861

Table 8: Classification Accuracy per DevEd Level, 3 Topmost Ranked Features (testing).

Results show RF, again, as the best-performing model with the same CA of 0.889. A slight improvement in the overall performance of the other three models, during this training phase is evidenced. The AUC scores also improved for all the models. This improvement, paired with the slight increase in the CA scores, reverses the order from NN>NB, in the training phase, to NB>NN, in the testing phase. The slightly worst results from this experiment, using the topmost ranked features, suggest that all features here used contribute in some way or to some degree to the classification process, even if some features correlate better with the overall classification.

Finally, Pearson coefficient calculations were performed. Results revealed that human ratings correlate poorly with the Accuplacer placement (Pearson r=0.172, n=98), even when the two raters agree on the essay skill level (Pearson r=0.301, n=27). These figures suggest that Accuplacer's placement cannot reliably correlate with human classification, at least based on the criteria explicitly elected (and allegedly used by this system), and using the data here collected (though the size of the sample is small).

In alignment with the objectives of this study, presented in Section 1, the experiments performed in this section aimed to pinpoint how Accuplacer's guidelines cannot be consistently applied by human raters. Eventually, the goal of this research is to support a more systematic placement of students by enhancing such an automatic classification system.

6. Conclusions and Future Work

This paper presented a first step toward assessing and selecting relevant features that could contribute to estimating or modeling the language proficiency of native English speakers in view of automatic DevEd placement.

The study highlighted the inadequacies of a widely-used automatic classification system – Accuplacer– emphasizing the importance of placements that truly reflect students' linguistic abilities. Proper placement supports students' literacy development and enhances higher education learning efficiency, making it not only a technical but also an ethical and economical necessity. Current systems, including Accuplacer, as evidenced in its manual (The College Board, 2022), offer limited linguistic feature definitions and lack transparency in their classification process details (rater profiles, participating institutions, and the sample size that is used for machine-learning training).

This study introduced clearer annotation guidelines, particularly for *Mechanical Conventions* and *Organization & Structure*, relating specific errors and paragraph structures to proficiency levels. It also revealed challenges in annotation due to its complexity and human inconsistency in rating. Out of all 49 essays assessed by at least two (2) raters, only 27 of them (55%) received the same overall skill classification (DevEd level 1 or 2). Despite low overall consistency scores, refining feature definitions led to better results in certain categories.

The study suggests the need for further precision in feature definitions and expanded classification guidelines. Future work includes exploring the integration of features automatically extracted from the texts using NLP-based tools like Coh-Metrix[9] (McNamara et al., 2006) and CTAP[10] (Chen and Meurers, 2016). These tools have played a crucial role in the analysis of linguistic complexity across various languages. It is expected to automatically extract relevant linguistic features to the DevEd setting that can enhance the classification process. Additionally, forthcoming research aims to leverage the Orange text mining tool with a larger annotated dataset, aiming to improve the reliability and accuracy of the placement process.

7. Ethical Considerations and Limitations

This study utilized a systematic sampling method, adhering to TCC's Institutional Review Board (IRB) protocols[11], which guaranteed ethical, fair, and equitable participant selection and protection. Approved by the IRB with the identifier #22-05, the research focused on educationally disadvantaged individuals, rigorously following IRB guidelines to both address and highlight the unique challenges faced by this group within an ethical framework.

8. Acknowledgments

This work was supported by Portuguese national funds through FCT (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837). We also extend our gratitude to the dedicated annotators who participated in this task for their meticulous work and innovative approach to help us advance our research.

9. Bibliographical References

Elisabeth A Barnett, Elizabeth Kopko, Dan Cullinan, and Clive R Belfield. 2020. Who should take college-level courses? Impact findings from an evaluation of a multiple measures assessment strategy. *Center for the Analysis of Postsecondary Readiness*.

Susan Bickerstaff, Elizabeth Kopko, Erika B Lewy, Julia Raufman, and Elizabeth Zachry Rutschow. 2021. Implementing and Scaling Multiple Measures Assessment in the Context of COVID-19. Research Brief. *Center for the Analysis of Postsecondary Readiness*.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Cohen. 1988. *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Maria Cormier and Susan Bickerstaff. 2019. Research on developmental education instruction for adult literacy learners. *The Wiley Handbook of Adult Literacy*, pages 541–561.

Miguel Da Corte and Jorge Baptista. 2024a. Classification of writing proficiency in developmental education. GitLab repository.

Miguel Da Corte and Jorge Baptista. 2024b. Guidelines to participate in a classification task assessing writing proficiency in deved courses. GitLab repository.

[9] http://141.225.61.35/CohMetrix2017/

[10] http://sifnos.sfs.uni-tuebingen.de/ctap/

[11] https://www.tulsacc.edu/

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

Jane Denison-Furness, Stacey Lee Donohue, Annemarie Hamlin, and Tony Russell. 2022. Welcome/Not Welcome: From Discouragement to Empowerment in the Writing Placement Process at Central Oregon Community College. In Jassica Nastal, Mya Poe, and Christie Toth, editors, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsceondary Education*, pages 107–127. The WAC Clearinghouse/University Press of Colorado.

Nia Dowell and Vitomir Kovanovic. 2022. Modeling educational discourse with natural language processing. *Education*, 64:82.

Deen Freelon. 2013. Recal OIR: ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, 8(1):10–16.

Holly Hassel and Joanne Baird Giordano. 2015. The blurry borders of college writing: Remediation and the assessment of student readiness. *College English*, 78(1):56–80.

Lisa Hilte, Walter Daelemans, and Reinhild Vandekerckhove. 2020. Lexical patterns in adolescents' online writing: the impact of age, gender, and education. *Written Communication*, 37(3):365–400.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2021. Ethics of AI in Education: Towards a Community-wide Framework. *International Journal of Artificial Intelligence in Education*, pages 1–23.

Sarah Hughes and Ruth Li. 2019. Affordances and limitations of the accuplacer automated writing placement tool. *Assessing Writing*, 41:72–75.

TK Koo and MY Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Amy Mazzariello, Elizabeth Ganga, and Nikki Edgecombe. 2018. Developmental education: An introduction for policymakers. *Education Commission of the States*, pages 1–12.

Danielle S McNamara, Yasuhiro Ozuru, Arthur C Graesser, and Max Louwerse. 2006. Validating CoH-Metrix. In *Proceedings of the 28th annual Conference of the Cognitive Science Society*, pages 573–578.

Jane S Nazzal, Carol Booth Olson, and Huy Q Chung. 2020. Differences in academic writing across four levels of community college composition courses. *Teaching English in the Two Year College*, 47(3):263–296.

Ajay Kumar Pal and Saurabh Pal. 2013. Classification model of prediction for placement of students. *International Journal of Modern Education and Computer Science*, 5(11):49.

Dolores Perin, Julia Raufman, and Hoori Santikian Kalamkarian. 2015. Developmental reading and English assessment in a researcher-practitioner partnership. Technical report, CCRC, Teachers College, Columbia University.

Kaśka Porayska-Pomsta and Gnanathusharan Rajendran. 2019. Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. *Artificial Intelligence and Inclusive Education: Speculative Futures and Emerging Practices*, pages 39–59.

V Ramesh, P Parkavi, and P Yasodha. 2011. Performance analysis of data mining techniques for placement chance prediction. *International Journal of Scientific & Engineering Research*, 2(8):1.

The College Board. 2018. *ACCUPLACER program manual*. The College Board New York.

The College Board. 2022. ACCUPLACER Program Manual. (online).

Ke Zhang and Ayse Begum Aslan. 2021. AI Technologies for Education: Recent Research & Future Directions. *Computers and Education: Artificial Intelligence*, 2:100025.

Mengxiao Zhu, Ou Lydia Liu, and Hee-Sun Lee. 2020. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143:103668.