MIGUEL DA CORTE

LANGUAGE ASSESSMENT AND PLACEMENT

LINGUISTIC AND COMPUTATIONAL APPROACHES IN DEVELOPMENTAL EDUCATION

UAlg
UNIVERSIDADE DO ALGARVE

Faculdade de Ciências Humanas e Sociais

2025

MIGUEL DA CORTE

LANGUAGE ASSESSMENT AND PLACEMENT

LINGUISTIC AND COMPUTATIONAL APPROACHES IN DEVELOPMENTAL EDUCATION

Doutoramento em Ciências da Linguagem

Trabalho efetuado sob a orientação do:
Professor Doutor Jorge Baptista

UAlg
UNIVERSIDADE DO ALGARVE

Faculdade de Ciências Humanas e Sociais

2025

Declaração de Autoria de Trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

O candidato:

_____

Assinatura do Candidato

Miguel Da Corte

Dedication

First and foremost, I thank God for opening the doors to this opportunity at the University of Algarve—for the wisdom, strength, and signs that guided me every step of this journey. To Him, I dedicate the words He placed in my heart: *força e acredite*.

To my incredible parents, thank you for always believing in me. Mom, you are the calm in my storms, the voice of wisdom and encouragement. I am who I am today because of you. Your countless sacrifices shaped the very foundation of my being. You have always given your children your best, and to me, you are the best. Thank you for instilling in me the importance of faith and the power of believing, especially when the path ahead felt uncertain. Dad, your presence at every step of this journey and your commitment to excellence have been my compass. Thank you for instilling in me the importance of giving my very best.

To my sister, my companion and friend, thank you for walking this path with me, for sharing your insight, your light, and your heart. Your presence reminded me that I was never alone. Thank you for helping me see the bright spots in every step, and for reminding me to remain humble, collected, and human, even in moments of pressure. Thank you for being a constant and grounding force in my life.

To my husband, my rock, my sounding board, one of my biggest supporters, and the keeper of my sanity during early-morning presentations and long nights of data. Your kindness, patience, and belief in my dreams carried me through. For the peace and the smiles you brought into my life, I cannot thank you enough. Thank you for always being by my side, selfless and ever ready to lend a hand.

To my extended family and dear friends, thank you for every prayer, every word of encouragement, and every moment you took to ask about my progress. Your support lifted me more than you know.

To my professors and mentors, thank you for your guidance, your challenges, and your respect. You helped me see the world, languages, and academia through a renewed lens. Thank you for helping me grow while valuing my voice.

And finally, to life itself, for the lessons, the detours, the resilience, and for leading me, at last, to this finish line.

Abstract

This thesis addresses the challenges of language assessment and placement in Developmental Education (DevEd) for native English (L1) speakers, stemming from the limitations of automated systems such as ACCUPLACER. At Tulsa Community College (TCC), where the study was conducted, writing proficiency is assessed through ACCUPLACER, and students not deemed college-ready are placed in DevEd courses to strengthen their skills before advancing to college-level coursework. This thesis draws on nine peer-reviewed papers organized into four thematic sections and integrates Corpus Linguistics, Machine Learning (ML), and Natural Language Processing (NLP) to address these limitations. *Section 1* lays the groundwork by mapping the linguistic landscape of developing writers through an annotation framework with 21 lexical, syntactic, and discourse features tailored to the DevEd context (DES). This framework is tested, refined, and applied to a newly created and publicly documented corpus of DevEd writing samples. The framework's adaptability is further demonstrated through a cross-linguistic extension to Spanish. Building on this foundation, *Section 2* integrates the 21 DES features with 436 linguistic features automatically extracted from NLP tools (COH-METRIX, CTAP) to create a richer, more comprehensive feature set. This set is used to train and test supervised ML models against a human-produced gold standard, evaluating whether these models can replicate, and potentially improve upon, ACCUPLACER's placement accuracy. The findings from these ML classification experiments lead to the questions of fairness and validity, which are addressed in *Section 3*. This section examines whether ACCUPLACER's classifications reflect students' actual writing performance and whether placement outcomes are equitable across different gender and race groups. Finally, *Section 4* investigates the use of Large Language Models (LLMs) as complementary tools for writing placement. Through prompt engineering and comparative evaluation, this section explores how LLMs can offer adaptive, linguistically sensitive, and more transparent alternatives to current automated classification systems. In the conclusion, this thesis synthesizes these findings, outlining key contributions, broader implications, and future research directions, all of which advocate for more linguistically informed placement practices.

Keywords: Developmental Education (DevEd), Writing Placement, Native (Writing) Language Proficiency, Machine Learning (ML), Natural Language Processing (NLP) tools, Large Language Models (LLMs)

Resumo

Esta tese aborda os desafios da avaliação de proficiência escrita e posterior colocação em programas de *Developmental Education* (DevEd) no sistema de ensino superior norte-americano. Este termo designa cursos de apoio pré-universitários destinados a colmatar lacunas académicas (p. ex., inglês, matemática), preparando-os para a frequência dos cursos superiores. Em concreto, o estudo analisa a produção escrita de falantes nativos de inglês (L1), centrando-se nas limitações de sistemas de avaliação e colocação automáticos. No Tulsa Community College (TCC), onde o estudo foi realizado, a proficiência em escrita é avaliada através de um teste feito com um desses sistemas, concretamente o ACCUPLACER. Os estudantes que são avaliados como não estando preparados para o ensino superior são colocados em cursos DevEd para reforçarem as suas competências antes de prosseguirem para unidades curriculares de nível superior. A tese organiza-se em quatro secções temáticas, apoiada em nove artigos, publicados em revistas e conferências internacionais com revisão por pares, e cruzando conceitos e ferramentas de Linguística de Corpus, Aprendizagem Automática (em inglês, *Machine Learning*, ML) e Processamento computacional de Linguagem Natural (*Natural Language Processing*, NLP) para responder às limitações identificadas. A *Secção 1* estabelece a base, mapeando o panorama linguístico de estudantes com proficiências de escrita em desenvolvimento, através de um quadro de anotação com 21 traços lexicais, sintáticos e discursivos, especificamente adaptados ao contexto DevEd (DES). Este quadro é testado, refinado e aplicado a um novo corpus de produções escritas de DevEd, documentado publicamente. A sua adaptabilidade é ainda demonstrada através de uma extensão da sua aplicação a um corpus semelhante do espanhol. Com base nesta fundação, a *Secção 2* integra as 21 características DES da Secção 1 com 436 características linguísticas extraídas automaticamente a partir de ferramentas de NLP (COH-METRIX, CTAP), criando um conjunto mais amplo e diversificado. Este conjunto serve para treinar e testar modelos de ML supervisionados para a tarefa de classificação e, depois, compará-los com um padrão de referência produzido por anotadores humanos, avaliando até que ponto estes modelos conseguem replicar — e eventualmente melhorar — a taxa de precisão do ACCUPLACER na colocação de estudantes. Os resultados destas experiências de classificação conduzem às questões da justiça e da equidade, exploradas na *Secção 3*, na qual se analisa se as classificações do ACCUPLACER refletem o desempenho real de escrita dos estudantes e se os resultados de colocação são equitativos entre diferentes grupos de género e raça. Por fim, a *Secção 4* investiga o uso de Modelos de Linguagem de Grande Escala (*Large Language Models*, LLMs) como ferramentas complementares de colocação em escrita. Através da engenharia de instruções (*prompt engineering*) e de uma avaliação comparativa, esta secção discute o potencial dos LLMs em oferecer alternativas adaptativas, linguisticamente sensíveis e mais transparentes face aos sistemas automáticos atualmente em uso. A conclusão sintetiza os resultados deste estudo, destacando as principais contribuições, as implicações mais amplas e direções futuras

de investigação, em defesa de práticas de avaliação de proficiência linguística e de colocação mais justas e informadas.

Palavras-chave: *Developmental Education* (DevEd), Colocação em Escrita, Proficiência na Escrita em Língua Materna, Aprendizagem Automática, Processamento de Linguagem Natural, Modelos de Linguagem de Grande Escala

Extended Abstract

This thesis addresses the challenges of language assessment and placement in Developmental Education (DevEd) for native English (L1) speakers, where English serves as the primary language of instruction and evaluation. These challenges stem from the limitations of automated classification systems, more precisely ACCUPLACER. At Tulsa Community College (TCC), where this study was conducted, students demonstrate writing proficiency through an entrance exam administered using ACCUPLACER. Students who are not deemed college-ready are placed in Developmental Education (DevEd) courses to strengthen their linguistic skills until they reach the proficiency required to participate in college-level classes. Among the limitations of these automated placement systems, placement accuracy is a salient issue. Drawing on nine published, peer-reviewed papers organized into four thematic sections, this thesis integrates Corpus Linguistics methodology with Machine Learning (ML) and Natural Language Processing (NLP) to address these limitations by increasing assessment transparency, improving placement accuracy, and ensuring that placement decisions more accurately reflect students' writing proficiency.

Section 1 centers on three papers that map the linguistic landscape of developing writers through an annotation scheme tailored to native language proficiency. Using a Corpus Linguistics approach, the first paper reports on a study that assembled a 100-essay corpus of DevEd student writing from TCC's placement exam database, classified by ACCUPLACER into DevEd Level 1 and DevEd Level 2. A pilot annotation produced a 21-feature DevEd-specific set (DES) spanning orthographic, grammatical, lexical, and discursive patterns, capturing both weaknesses and strengths in students' writing. This framework served as the basis for annotator guidelines, training, and quality checks on annotations. The second paper builds on these results by focusing on Multiword Expressions (MWE). A custom annotation scheme was developed to categorize diverse MWE types and applied to all instances in the original English corpus of 100 texts. Using this annotated corpus, ML models available via the data-mining tool ORANGE were trained and tested to examine whether the presence and types of MWE carried predictive value in placement decisions. The third paper extended the MWE annotation framework to a Spanish-language corpus of texts produced by native and non-native speakers of such language in a Spanish developmental writing course at TCC. This cross-linguistic application assessed whether MWE functioned as reliable proficiency indicators beyond English, confirming patterns from the English (L1) corpus. ML experiments in ORANGE further tested the predictive value of specific MWE types and explored how these findings might inform pedagogical practice.

Section 2 builds on the annotated English (L1) corpus and the DES feature framework, shifting the focus from annotation to evaluation, through three papers. The first paper reports on a study that used the original 100-text corpus, classified by ACCUPLACER, but now it is also rated by skill level (DevEd Level 1 or DevEd Level 2) by trained human evaluators according to ACCUPLACER's descriptors. These descriptors do not disclose the specific linguistic features included within each category or the weighting of these features in the assessment process.

Inter-rater reliability was calculated to measure agreement in the skill-level classifications and to determine if refinements to the linguistic descriptors were needed. Discrepancies were settled by a third rater, establishing the first step toward a validated gold standard for DevEd writing classification. The second study reports on a study that expanded the linguistic analysis by combining the initial 21 DES features framework, devised in Section 1, with automatically extracted metrics from two NLP tools: Coh-Metrix and CTAP, creating a 457-feature set that combines pedagogical insight with computational precision. ML experiments were conducted on the same 100-text corpus, using both full-length texts and segmented texts (each split into 100-word segments) to examine the impact of text length on placement accuracy. These experiments pursued two main goals: (i) identifying the most predictive linguistic features and (ii) determining which algorithms performed best for writing placement. Four experimental designs provided complementary insights: (1) establishing baselines for each feature set independently; (2) testing the top features from each source ranked by Information Gain and Chi-square scoring methods; (3) applying a leave-one-cluster-out method to measure the contribution of lexical, semantic, and discourse clusters; and (4) combining features from all sources, ranked in groups of ten until ML performance plateaued. The third paper reports on a study that aimed to combine interpretable, human-devised linguistic markers with automatically extracted NLP-based features to improve placement accuracy. To support this goal, the original 100-text corpus was expanded to 300 essays. All 200 new essays, randomly selected from the same entrance exam database, were annotated by the same two human raters, who also assigned a skill level to each text. Using the Information Gain scoring method in Orange, the original 457-feature set was refined to 445 features to assess (i) whether the previously top-ranked features continued to demonstrate strong predictive power for placement outcomes and (ii) whether the updated feature set led to higher classification accuracy scores in the ML models. The ML experiments, conducted in Orange, were organized into two scenarios: (1) using the full corpus with classification levels assigned automatically by Accuplacer, and (2) using the same corpus with levels assigned by human annotators. In both scenarios, experiments began with the refined feature set. Then, features were grouped by Information Gain rankings and incrementally added in bundles of ten until the ML models' performance plateaued, mirroring the approach used in the second study of this section.

Section 3 presents two papers that examined whether Accuplacer's classifications reflect students' actual writing performance and whether placement outcomes are equitable across the sociodemographics of the participants in this study. The first paper builds on a study where human raters applied the refined Accuplacer descriptors from Section 2 to the original corpus, assigning each text its corresponding skill level. A third rater resolved the skill-level disagreements, helping create the gold standard used in subsequent analyses. This gold standard served as the basis to inform classical ML experiments in Orange with two objectives: (i) to identify which linguistic descriptors were most predictive of placement levels and how they influence classification accuracy, and (ii) to measure Accuplacer's accuracy against that of human anno-

tators applying the same linguistic criteria. The second paper builds on a study that examined how the linguistic features explored throughout this investigation, together with placement outcomes from both ACCUPLACER and human raters, intersected with sociodemographic variables (gender and race) in a subset of 210 essays. To this end, three statistical tests were employed: (1) ANOVA to explore differences in feature use across gender and racial groups; (2) Tukey's HSD to identify where within those groups the differences occurred; and (3) Chi-square to assess associations between sociodemographic variables and placement outcomes.

Section 4 presents one paper that explores whether Large Language Models (LLMs) can serve as complementary placement tools to ACCUPLACER. Using an expanded 450-essay corpus, models including OpenAI ChatGPT, Claude, and Gemini were evaluated against ACCUPLACER and the human gold standard under two strategies: (1) a one-step classification, in which the LLMs classified all texts in one run into three levels (DevEd Level 1, DevEd Level 2, or College-level), and (2) a two-step classification, in which the LLMs classified all texts first into College and DevEd and, then, in a second run, they classified the DevEd texts only into Levels 1 or 2. Four prompting conditions, namely, (a) Zero-shot, (b) Few-shot, (c) Enhanced, and (d) Enhanced+, were tested to determine whether detailed linguistic markers improved the classification precision of the algorithms.

This thesis concludes by synthesizing the contributions of these nine studies, outlining broader implications and future research directions, and advocating for placement practices that integrate interpretable linguistic features with NLP, ML, and LLMs to better support student success in higher education.

Resumo alargado

Esta tese aborda os desafios da avaliação da proficiência linguística e da colocação de estudantes em programas de *Developmental Education* (DevEd) — termo utilizado no ensino superior dos Estados Unidos norte-americanos para designar cursos de apoio pré-universitários destinados a colmatar lacunas académicas em inglês, noutras disciplinas como a matemática — dirigidos a falantes nativos de inglês (L1), onde esta língua é também o principal veículo de instrução e avaliação. Tais desafios resultam das limitações dos sistemas automáticos de classificação. No Tulsa Community College (TCC), onde este estudo foi realizado, os estudantes demonstram a sua proficiência escrita em inglês através de um exame de admissão administrado por um destes sistemas automáticos, concretamente, o ACCUPLACER. Aqueles que não são considerados preparados para o ensino superior são colocados em cursos DevEd para assim reforçarem as suas competências até atingirem o nível exigido para prosseguir para unidades curriculares do nível do ensino superior (College). Entre as limitações dos sistemas de colocação automática, a precisão da colocação constitui uma questão central. Esta tese tem por base nove artigos publicados em revistas e conferências internacionais com revisão por pares, organizados em quatro secções temáticas. Nela se integram metodologias da Linguística de Corpus, de Processamento de Linguagem Natural (NLP) e de Aprendizagem Automática (ML) com o objetivo de responder a estas limitações, aumentando a transparência da avaliação, melhorando a precisão da colocação e assegurando que as decisões reflitam de forma mais rigorosa a real proficiência escrita dos estudantes em língua inglesa.

A Secção 1 reúne três artigos que mapeiam o panorama linguístico de estudantes enquanto escritores em desenvolvimento, através de um quadro de anotação adaptado à proficiência em língua materna. O primeiro artigo apresenta a construção de um corpus de 100 redações de estudantes de DevEd, retiradas da base de dados do TCC e classificadas pelo ACCUPLACER nos níveis DevEd 1 e 2. Uma anotação-piloto produziu um conjunto específico de 21 características de DevEd (DES), abrangendo padrões ortográficos, gramaticais, lexicais e discursivos, captando tanto fragilidades como pontos fortes. Este quadro serviu de base para a elaboração de diretrizes de anotação, a formação de anotadores e o controlo da qualidade da anotação. O segundo artigo desenvolve estes resultados centrando-se em expressões multipalavra (ing., *multiword expressions*, MWE), através de um esquema de anotação específico, com o qual se categorizaram diferentes tipos de MWE em todo o corpus. Este recurso foi utilizado para treinar e testar diversos modelos de ML na plataforma de mineração de dados (ing., *data mining*) ORANGE. O objetivo consistiu em avaliar se a presença e o tipo de MWE tinham valor preditivo nas decisões de colocação. O terceiro artigo estendeu o quadro de anotação de MWE a um corpus de textos em espanhol, produzidos por falantes nativos e não nativos num curso de desenvolvimento de escrita (em espanhol), realizado no TCC. Procurou-se verificar se estas expressões funcionavam como indicadores fiáveis de proficiência para além do inglês.As experiências de ML no ORANGE

testaram ainda o valor preditivo de categorias específicas de MWE e exploraram as respetivas implicações pedagógicas.

A Secção 2 desloca o foco da anotação para a avaliação, com três artigos baseados no corpus de inglês (L1) e no respetivo quadro de características. O primeiro artigo apresenta um estudo em que os mesmos 100 textos iniciais foram também avaliados por dois anotadores humanos, seguindo os descritores propostos pelo ACCUPLACER. Como estes descritores não especificam as características linguísticas em causa nem o peso atribuído a cada uma, foi calculado o acordo interavaliador e, no caso de haver discrepâncias, estas resolvidas por um terceiro avaliador, o que constituiu um primeiro passo para estabelecer um padrão de referência validado para a classificação da escrita em DevEd. O segundo estudo ampliou a análise ao combinar as 21 características DES da Secção 1 com 436 métricas extraídas automaticamente de ferramentas de NLP (COH-METRIX, CTAP), criando um conjunto de 457 características. As experiências de ML foram conduzidas com o corpus de 100 textos, recorrendo tanto aos textos integrais (de tamanhos variáveis) como aos textos segmentados em blocos de 100 palavras, a fim de analisar o impacto da extensão textual na precisão da colocação. Quatro desenhos experimentais forneceram perspetivas complementares: (1) estabelecer linhas de base (ing., *baselines*) para cada conjunto de características isoladamente; (2) testar as principais características de cada fonte (COH-METRIX, CTAP e DES), classificadas pelos índices de Ganho de Informação (ing., *Information Gain*) e Qui-quadrado, disponíveis na plataforma ORANGE; (3) aplicar o método de validação cruzada *leave-one-out* (LOO) para determinar a contribuição de agrupamentos de características lexicais, semânticos e discursivos; e (4) combinar as características mais informativas de todas as fontes em conjuntos progressivos de dez até a estabilização do desempenho. O terceiro artigo procurou combinar marcadores linguísticos interpretáveis, concebidos por humanos, com características extraídas automaticamente por NLP. Para isso, o corpus foi expandido de 100 para 300 redações, e as 200 novas amostras foram anotadas pelos mesmos avaliadores humanos que anotaram o conjunto inicial de 100 textos. Com recurso ao índice de Ganho de Informação, o conjunto de 457 características foi refinado para 445, tendo-se avaliado se as características mais relevantes mantinham o poder preditivo e se o conjunto atualizado aumentava a precisão dos modelos de ML. As experiências foram ainda organizadas em dois cenários: (1) corpus classificado automaticamente pelo ACCUPLACER; e (2) corpus classificado por anotadores humanos, replicando a abordagem incremental do estudo anterior.

A Secção 3 apresenta dois artigos dedicados à validade e à equidade da colocação. O primeiro estudou se as classificações do ACCUPLACER refletem efetivamente o desempenho e a proficiência na escrita dos estudantes, aplicando os descritores refinados da Secção 2 ao corpus e comparando a precisão das classificações automáticas e humanas em diferentes experiências de ML. O segundo artigo analisou se variáveis sociodemográficas, nomeadamente, o género e a raça, influenciavam os resultados da colocação. Num subconjunto de 210 redações, representativo dos estudantes do TCC, foram aplicados três testes estatísticos: (1) ANOVA (para diferenças

entre grupos), (2) Tukey's HSD (para localizar diferenças específicas) e (3) Qui-quadrado (para determinar as associações entre variáveis sociodemográficas e resultados de colocação).

A Secção 4 explora em que medida Modelos de Linguagem de Grande Escala (LLMs) podem servir como ferramentas complementares de colocação. Com um corpus expandido de 450 redações, balanceado por níveis DevEd e universitário, alguns modelos, como o ChatGPT, o Claude e o Gemini, foram comparados com a classificação produzida pelo ACCUPLACER e com a classificação de referência, produzida por avaliadores humanos. Adotaram-se duas estratégias de classificação: (1) classificação multinível num único passo (três níveis: DevEd 1, DevEd 2, Ensino Superior); e (2) classificação em dois passos, separando primeiro os níveis DevEd do Ensino Superior, e, depois, subdividindo os níveis DevEd em Nível 1 e Nível 2. Foram testadas quatro condições de instrução (*prompting*): *Zero-shot* (sem exemplos), *Few-shot* (com poucos exemplos), *Enhanced* (instruções detalhadas) e *Enhanced+* (instruções detalhadas com reforço adicional). O objetivo desta estratégia foi avaliar se a inclusão de marcadores linguísticos detalhados permitia aumentar a precisão de classificação.

Esta tese conclui com uma síntese das principais contribuições dos estudos apresentados, destacando as suas implicações mais amplas e indicando direções futuras de investigação, defendendo práticas de colocação que combinem características linguísticas interpretáveis, com recurso a Processamento Computacional de Linguagem Natural (NLP), Aprendizagem Automática (ML) e Modelos de Língua de Grande Escala (LLM), de modo a apoiar de forma mais justa e eficaz o sucesso académico dos estudantes no ensino superior.

Note to Reader

This thesis adopts an interdisciplinary approach, bridging the fields of Applied Linguistics and Corpus Linguistics, and represents the culmination of three years of intensive reading and research. Presented as a compilation of papers[1], it brings together nine papers organized into four thematic sections. Eight papers were written in English, my second language, and one in Spanish, my first language, as required by the journal of publication.

Together, these papers were deemed sufficient to demonstrate the rigor and soundness of this investigation into language assessment and placement in the context of DevEd. They were double-blind, peer-reviewed, published, and/or presented at international, competitive conferences held in the Czech Republic, France, Italy, Spain, and, last but not least, Portugal, a country I also call home, given my Portuguese heritage.

Although the nine papers are not presented in chronological order, their organization follows a carefully structured thematic sequence, with each paper including its own dedicated literature review. Each section builds upon the previous one, reflecting the methodological and analytical progression used to: (i) explore the overarching research questions; and (ii) develop the linguistic and pedagogical implications of this work. To clarify the structure and purpose of each section, an introductory overview precedes the presentation of the papers.

Finally, each paper adhered to the formatting and editorial guidelines of the journal or conference in which it was published and is included here in its original form. Compiling them according to the thesis formatting requirements would result in a document exceeding the 250-page limit suggested in the regulation cited above. Due to my institutional affiliation and the supervised nature of the research, my advisor is listed as a co-author on all papers that I have authored.

[1] Regulation No. 114/2023, of January 23, *Diário da República*, Series II, No. 16, Art. 27.c).i); Annex 1, No. 1.9.

Contents

List of Abbreviations

This table presents the most commonly used terms throughout this document, in alphabetical order, along with their corresponding abbreviations.

Terms	Abbreviations
Adaptive Boosting	AdaBoost
American Association of Community Colleges	AACC
Argumentation with Examples	EXAMPLE
Argumentation with Reasoning	REASON
Area Under the Curve	AUC
Artificial Intelligence	AI
Classification Accuracy	CA
CN2 Rule Induction	CN2
College-Level Placement	CA
Common Text Analysis Platform	CTAP
Decision Tree	DT
DevEd-Specific (features)	DES
Developmental Education	DevEd
Developmental Education – Level 1	DevEd Level 1
Developmental Education – Level 2	DevEd Level 2
First Language	L1
Gradient Boosting	GB
Grapheme-level Errors	ORT
Information Gain	IG
Institutional Review Board	IRB
Intraclass Correlation Coefficient	ICC
k-Nearest Neighbors	kNN
Krippendorff's alpha	K-alpha
Large Language Model	LLM
Lexical Precision	PRECISION
Logistic Regression	LR
Machine Learning	ML

Continued on next page

Continued from previous page

Terms	**Abbreviations**
Measure of Textual Lexical Diversity	MTLD
Multiword Expression	MWE
Naïve Bayes	NB
National Center for Education Statistics	NCES
Natural Language Processing	NLP
Neural Network	NN
New General Service List	NGSL
Oklahoma State Regents for Higher Education	OSRHE
Organization for Economic Co-operation and Development	OECD
Parts of Speech	POS
Purpose and Focus	PF
Random Forest	RF
Representativeness Index	RI
Second Language	L2
Sentence Variety and Style	SVS
Support Vector Machine	SVM
Test of English as a Foreign Language	TOEFL
Type-Token Ratio	TTR
Tulsa Community College	TCC
Tukey's Honestly Significant Difference	Tukey's (HSD)
The United States of America	U.S.
Verb Agreement	VAGREE

Introduction

This thesis investigates the challenges of language assessment and placement in a community college in the United States of America (U.S.), with a specific focus on Developmental Education (DevEd) programs for native English speakers (L1). These challenges stem from the limitations of existing automated classification systems, particularly in accurately capturing and assessing students' linguistic (writing) proficiency. Such limitations can lead to misplacement within course sequences, delaying students' progress in their chosen academic programs.

This introduction begins by contextualizing the role of community colleges in providing academic and linguistic support to students through Developmental Education (DevEd) programs. It then offers a demographic overview of the institution where the study was conducted, highlighting key characteristics of the student body and introducing the study participants. From there, the focus shifts to the central concern of this thesis: *the assessment of participants' English (L1) writing skills and the placement outcomes that result from such assessment*. To ground this work, the introduction next outlines the linguistic framework guiding the investigation, including the corpus collection and annotation processes. This framework is followed by the computational techniques employed (Natural Language Processing and Machine Learning) and the four research questions set forth for this study. This introduction closes with a roadmap of the sections that shape this thesis.

Community Colleges and Developmental Education (DevEd)

Community colleges are higher education institutions tasked with delivering the first two years of a Bachelor's degree, awarding certificates in specialized trades, and providing academic preparation for occupational credentials, according to the U.S. Department of Education[2]. They serve an academically diverse population, including recent high school graduates and adults returning to education, many of whom do not yet meet the literacy standards typically associated with college readiness (Kosiewicz et al. (2024)). The National Center for Education Statistics (NCES)[3] defines literacy both conceptually, as "the ability to use printed and written information to function in society, achieve one's goals, and develop one's knowledge and potential," and operationally, as "the successful use of printed material," a task that depends on both word-level reading skills and higher-level literacy skills. When there is a potential gap between these skills and the proficiency expected for college-level work, particularly in reading and writing, it can directly influence whether a student begins in credit-bearing coursework or DevEd[4] courses (Holschuh (2019); MacArthur (2023)).

[2] https://www.ed.gov/higher-education (All URLs in this thesis were checked on September 11, 2025.)

[3] https://nces.ed.gov/naal/

[4] The terms *developmental* and *remedial*, as applied to education or courses, are used interchangeably throughout this document.

DevEd courses, often offered at community colleges, are designed to close this gap, providing targeted instruction in reading, writing, and mathematics to prepare students for the demands of credit-bearing coursework (Mazzariello et al. (2018); Cormier and Bickerstaff (2019); Bickerstaff et al. (2022)). While all three skill areas are important, this thesis centers on writing proficiency because it serves a dual role: a measure of academic readiness (Gere (2019)) and a gatekeeper to degree-bearing courses. In practice, proficient writing skills enable students to communicate effectively, engage with discipline-specific content, and progress toward an academic credential. On the contrary, non-proficient writing can delay graduation, increase costs, and potentially limit opportunities post-graduation. These costs include not only the financial burden of extended tuition, course materials, and institutional expenses such as faculty time and instructional resources, but also the loss of opportunities for workforce entry or advanced study.

Determining who requires DevEd support begins with the completion of an entrance exam often administered through ACCUPLACER[5], a test developed by the College Board (The College Board (2022)) and widely used at U.S. community colleges. As part of this exam, students respond to one of several preselected essay prompts under standardized testing conditions. The prompts are designed to elicit responses on a range of argumentative or reflective topics, such as *Is History Valuable*, *Practical Skills*, *Acquisition of Money*, and *Differences Among People*. All essays are produced in a controlled, proctored environment at the administering institution's testing center, without access to the Internet or editing tools. Each student's written response is then automatically assessed by ACCUPLACER, which produces a score that, in accordance with each institution's guidelines, determines whether a student meets the writing proficiency standard for college-level coursework or requires DevEd support.

At Tulsa Community College (TCC)[6], a public community college in the State of Oklahoma, United States of America (U.S.), and the site of this study, students are classified into one of three writing proficiency levels following completion of the ACCUPLACER placement test. The definitions of the three placement levels outlined below are adapted from official course descriptions[7], archived in the institution's academic catalog[8]:

- **DevEd Level 1**: for support with frequent grammar, spelling, and punctuation issues, and lacking cohesion;
- **DevEd Level 2**: for support with some structural improvements still needing targeted support;
- **College Level**: no DevEd support needed; skills show academic-level writing with minimal errors.

[5] https://accuplacer.collegeboard.org/

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

[7] At the time this study was conducted, the courses associated with each level were: DevEd Level 1 — Writing Foundations I; DevEd Level 2 — Writing Foundations II; and College level — English Composition I.

[8] https://catalog.tulsacc.edu

Before taking specific required college-level courses, students at TCC must demonstrate writing proficiency, ideally within their first year of a college career, according to the Oklahoma State Regents for Higher Education (OSRHE)[9]. For faculty and policymakers, understanding how placement decisions are made and who gets assigned to each level is essential, not only from an institutional standpoint but also from a pedagogical and linguistic perspective. Community colleges, therefore, carry an academic and social responsibility to support students in developing the language skills necessary for degree completion and graduation. At the same time, the field of linguistics offers the analytical tools and frameworks required to assess (and strengthen) those skills effectively.

Because placement decisions occur within an institutional context, and, for this study, it is TCC, a snapshot of its demographics during the research period (2021–2022 and 2023–2024) helps contextualize the potential impact of these decisions.

Institutional Demographics

According to data from TCC's Institutional Research, Reporting, and Analytics department[10], in 2023–2024, the institution served a population of 14,538 students. By race, the majority of the student population self-identified as White (nearly 50%), followed by Hispanic/Latino (16%), Black or African American (8%), and American Indian (7%) students. The remaining students either identified with more than one race or chose not to disclose their racial background at the time of application. By gender, 64% identified as female and 36% as male.

The 2023–2024 student population at TCC was predominantly part-time (69%), with only 31% enrolled full-time. The average student age was 23, with 57% aged 24 or younger, and nearly one in four (24%) identifying as first-generation college students. That same year, almost half of all students were enrolled in at least one DevEd course, primarily in reading and/or writing. After placement, these students enrolled in their recommended DevEd course(s). While a slight majority (56%) passed their courses with a grade of C or better, a substantial proportion (44%) either failed, withdrew, or stopped attending. These outcomes contribute to challenges in academic preparedness, delay program completion, and impose financial burdens on both students and the institution.

TCC's reality reflects a shared concern about academic preparedness that extends well beyond the institution, resonating with trends nationwide. According to the Organization for Economic Co-operation and Development (OECD)[11], 28% of adults in the U.S. demonstrate low literacy skills (OECD (2023)). In Oklahoma, specifically, $\approx$ 20% of adults have limited literacy, and $\approx$ 19% do not hold a high school diploma, placing the state 29$^{\text{th}}$ out of 50 in adult literacy rankings[12]. Addressing these challenges requires institutions not only to be ready

[9] https://okhighered.org/state-system/policy-procedures/
[10] https://www.tulsacc.edu/about-tcc/institutional-research
[11] https://www.oecd.org/en
[12] https://map.barbarabush.org/state-cards/

but also equipped with the tools to support students through programs like DevEd, given that effective communication is both an academic requirement and a life skill. Providing adequate support depends on accurate placement, ensuring students are enrolled in courses that match their skill level and allow them to progress effectively. This effort, in turn, calls for placement systems to be grounded in linguistic evidence, ensuring not only proper placement but also the provision of actionable feedback that faculty can use to tailor instruction until students reach the proficiency required for college-level success.

With the institutional demographics explained, it is pertinent to focus next on the students whose writing directly informed this study. These students completed the ACCUPLACER writing assessment, and their written productions form the corpus for this investigation, providing the foundation for a linguistic analysis aimed at defining what developmental writing looks like.

Study's Participants

This investigation focuses on students who intended to enroll in an academic program at TCC and completed the ACCUPLACER writing assessment as part of their entrance exam during the 2021–2022 and 2023–2024 academic years. To ensure ethical rigor and protect participants' rights, this thesis adhered to TCC's Institutional Review Board (IRB) protocols[13], under identifier #22-05, with special attention given to students classified as educationally disadvantaged, which is a designation that applies to those in DevEd.

Across the research period, approximately 2,003 students[14] (unduplicated count) completed and submitted their (typed) essay responses under proctored conditions. These essays, as previously mentioned, were automatically assessed and classified by ACCUPLACER according to three proficiency levels. Of these 2,003 students, 1,384 (69%) reported English as their native language (L1), forming the subset used in this thesis. Within this English (L1) group, the gender and racial representativeness[15] of this subsample compared to the overall institution's population was determined by a Representativeness Index (RI) calculations[16]. By gender, 62% identified as female and 37% as male. Based on this demographic variable, the sample is highly representative of the population (both RIs above 0.97). By race, 41% self-identified as White, followed by 15% Black or African American and 10% American Indian. Within this distribution, White students were closely represented (RI = 0.85), American Indian students were moderately represented (RI = 0.57), and Black or African American students were somewhat more represented in the sample compared to the population (RI = 0.13).

[13] TCC's Institutional Review Board webpage

[14] Because data collection occurred at defined intervals throughout 2021–2024, this sample reflects the essays available during those collection windows.

[15] Representativeness, as defined by Temnikova et al. (2014, p. 1714), "is the extent to which a corpus or other language sample reflects the language from which it is sampled."

[16] The RI was calculated by comparing the proportion of each race and gender category in the sample with its proportion in the population, expressed as one minus their relative difference.

Having outlined the broader context of community colleges and DevEd, and described the population of the institution where this study was conducted, the next section provides a more detailed explanation of why this thesis focuses specifically on English (L1) writing.

English (L1) Writing: Focus and Assessment

Because writing proficiency functions as one of the required standards for course enrollment at TCC, it serves as the medium through which this study examines the question of assessment. The decision to focus on English (L1) speakers aligns with this context, where English is the primary language of instruction and evaluation. It also establishes the foundation for this thesis's guiding question: *How can the writing of community college students be more accurately and equitably assessed, based on their written productions, to determine whether they require remediation or are ready for college-level coursework*? While reading remains foundational to academic development, writing entails a more integrated and demanding measure of preparedness (Gere (2019)).

At the college level, students are expected to produce well-developed texts that demonstrate control over sentence variety, tone, and vocabulary, and to support ideas through quoting, paraphrasing, and summarizing, which are essential skills across nearly every college discipline (Kim et al. (2025)). These expectations align closely with broader definitions of proficient writing, such as those articulated in the Test of English as a Foreign Language (TOEFL)[17], which emphasize coherence, organization, and grammatical precision. Together, these academic expectations and standardized descriptors highlight why writing proficiency functions as a gatekeeper for academic progression, making it a particularly revealing area for assessment and support (Kim et al. (2024)).

At TCC, participation in college-level coursework is determined in part by writing proficiency, which, as noted earlier, is assessed through the ACCUPLACER placement exam. While designed to facilitate placement into either college-level or DevEd courses, ACCUPLACER relies on a single, timed essay response and functions as a "black-box" algorithm: the criteria it applies and the linguistic features it considers are neither clearly documented nor publicly accessible. As a result, placement decisions lack the granularity needed to identify the lexical, syntactic, and discursive patterns that distinguish developmental from proficient writing. Moreover, students are left with unanswered questions about what college-readiness standards entail (Hirsch et al. (2024)), how their writing was evaluated, and what steps they could take to improve it (Murray and Sharpling (2019)). This lack of information generates skepticism about the transparency of the placement process, limiting students' ability to make informed decisions about their academic readiness.

The skepticism here mentioned is reinforced by evidence from the literature showing that automated placement systems in higher education misclassify roughly one in three students

[17] https://www.ets.org/toefl.html

(Cormier and Bickerstaff (2019); Ganga and Mazzariello (2019)), placing them into courses that do not align with their actual writing skills. This high rate of misplacement underscores the pressing need for placement systems that can more accurately capture how students write at the point of entry into college (Ramesh et al. (2011); Pal and Pal (2013); Zhang and Aslan (2021)). At TCC, this projected one-third misplacement rate could potentially translate into approximately 4,845 students[18] who might be inaccurately placed through the ACCUPLACER exam, either into courses that underestimate or overestimate their writing abilities. Such misplacements could skew the composition of students in DevEd programs, leading to a potential misrepresentation of linguistically underprepared students (Bickerstaff et al. (2022); East and Slomp (2024)). When this occurs, the resulting inaccuracies disrupt students' academic trajectories and carry implications for institutional costs (e.g., prolonged tuition and instructional resources) and instructional efficiency (Hughes and Li (2019); Link and Koltovskaia (2023); Edgecombe and Weiss (2024)), as well as student motivation and engagement.

Within this context, the importance of accurate placement becomes even more pronounced, especially at this critical entry point. It increases the likelihood that students will be placed into appropriate coursework and graduate on time. Many DevEd students, in particular, express concern about the extent of their vocabulary and their ability to produce structured, coherent writing. Their focus on lexical richness, especially as it relates to active vocabulary size, reflects a growing awareness that vocabulary control is foundational to academic success (Hilte et al. (2020)). Yet under a placement exam restricted to a single, timed essay, where neither the criteria applied nor the linguistic features considered are disclosed, students cannot know whether the writing issues they are concerned about are actually reflected in their written productions, leaving them without meaningful insight into their performance.

Because writing proficiency is central to academic readiness, this study responds to these issues by analyzing students' written productions to identify salient linguistic patterns that enable a more comprehensive assessment of vulnerable populations (Gilman et al. (2019); Nastal (2019); Nastal and Messer (2025)). Particular attention is given to DevEd students, where language proficiency determines access to higher education and instruction plays a vital role in supporting linguistically underprepared learners (Darkenwald-DeCola (2021); Gilquin and Granger (2022)). Identifying these patterns, such as how students use vocabulary within a sentence or construct meaning across sentences, offers a more systematic and measurable way to describe native language proficiency (Khan (2019); Kosiewicz et al. (2024)).

To ensure that such analysis is systematic and anchored in the existing literature, it is necessary to develop a framework that links theory to measurable indicators of writing proficiency. The following section introduces the linguistic and computational framework that grounds this thesis.

[18]Figure calculated based on TCC's reported enrollment of 14,538 students in the 2023–2024 academic year.

Linguistic and Computational Framework

This thesis is anchored in a linguistic and computational framework that integrates insights from Applied Linguistics and Corpus Linguistics. These approaches provide systematic methods for tracing and capturing the lexical, syntactic, and discursive patterns that can better define writing readiness. In terms of English and its patterned nature, Hornby (1954) laid early groundwork in categorizing verb usage by syntactic and functional patterns, emphasizing the importance of vocabulary knowledge alongside grammatical structure. Further support for a pattern-based approach comes from Hunston and Francis (2000), who classified verbs and nouns into recurring grammatical structures, providing a corpus-driven model for syntactic analysis. This notion of patterned language aligns with the work of Willis (2003) and Hanks (2008), who explained the role of frequent word combinations in shaping everyday communication. Building on this foundation, Schmitt and Schmitt (2020) highlighted the importance of multiword units in native speaker discourse, reinforcing the need to examine not only isolated vocabulary items, but also the formulaic language that students use.

These pattern-based approaches align with findings from recent annotated learner corpus research (Götz and Granger (2024)), which highlights the instructional benefits of exposing students to expert writing as a means of (1) encouraging reflection on interlanguage features (e.g., misuse, overuse, underuse); and (2) engaging in revision practices that strengthen writing proficiency over time. The combinations of all of these approaches led to a DevEd framework with features that extends across four main linguistic pattern domains:

- **Orthographic features** such as punctuation, spelling, and contractions that often indicate early-stage writing development.
- **Grammatical features** commonly targeted in learner corpora, including verb tense, agreement, and word omissions or additions.
- **Lexical and semantic features** that introduce clarity to the meaning of the sentences in a text, such as word choice and use of connectives.
- **Discursive features** reflecting a student's ability to structure extended arguments, express abstract thinking, and engage in academic-level discourse, all of which are essential skills for success in college-level writing.

Without a clear linguistic profile of how DevEd students write, it becomes difficult to determine whether current placement mechanisms, course sequences, and support structures are pedagogically justified. This ambiguity raises broader concerns about instructional efficiency, student progression, and the costs (not only economic, but also in terms of time and effort on the part of faculty) associated with misplacement. As a first step toward establishing such a profile, it was necessary to assemble a dedicated corpus of DevEd writing that could capture the actual linguistic patterns of community college students.

Before assembling the corpus for this research, and guided by a Corpus Linguistics methodology (McEnery et al. (2006); Xiao (2010); Durrant (2022)) to examine how linguistic patterns emerge in authentic learner data, an extensive review of existing English-language corpora was conducted to determine whether a resource that could capture the linguistic characteristics of community college DevEd writers already existed. While well-established corpora, such as the Corpus of Contemporary American English (COCA)[19] (Davies (2008)) and OntoNotes[20] (Pradhan et al. (2007)) include general and academic uses of English, the corpus needed for this thesis had to reflect more closely the writing patterns of community college students at the onset of their academic journey. To meet this need, a stratified random sample of 300 essays was drawn from the same TCC exam database that housed the 1,384 English (L1) responses produced during the research period, as mentioned in the Study's Participants section. Each text was anonymized and assigned a unique text identification number, in accordance with IRB protocols.

The 300-essay sample[21] was purposefully balanced by placement level (150 from DevEd Level 1; and 150 from DevEd Level 2) to facilitate detailed linguistic annotation and comparative analysis, ensuring equal representation across DevEd levels. This sample size, which balances statistical rigor with the practical constraints of manual annotation, serves as the corpus used in this research[22]. This corpus served as the basis for multiple studies within this thesis, beginning with its manual annotation using the linguistic framework introduced earlier.

These annotated texts served both as a reference for human-identified linguistic patterns and as input for a series of computational experiments that leveraged Natural Language Processing (NLP) and Machine Learning (ML) techniques. In this way, the corpus functioned not merely as a dataset but as the empirical ground on which the central guiding question of this thesis, *how to more accurately and equitably assess the writing of community college students*, could be addressed. Previous investigations (Abba et al. (2018); Leal et al. (2023)) focused on using NLP tools like Coh-Metrix[23] (McNamara and Graesser (2011)). In contrast, others (Okinina et al. (2020); Akef et al. (2024)) focused on the Common Text Analysis Platform (CTAP)[24] (Chen and Meurers (2016)), showing the value of features such as lexical diversity, phrase density, and cohesion in modeling student writing development across varied populations. By integrating manually annotated, novel, DevEd-inspired, specific linguistic features with automatically extracted ones, this thesis extends the state-of-the-art to investigate whether such a combination can yield more accurate and equitable assessments of student writing proficiency. This dual approach not only captures the linguistic patterns visible to human raters

[19] https://www.english-corpora.org/coca/

[20] https://catalog.ldc.upenn.edu/LDC2013T19

[21] As an illustration, **Appendix A** presents two sample essay responses from the corpus, each accompanied by their automatic classification score and feedback (DevEd Level 1 and DevEd Level 2) issued by Accuplacer.

[22] In accordance with institutional guidelines (IRB), full datasets will be released to the scientific community following the official presentation of this thesis, though selected sets are included in this document.

[23] https://soletlab.asu.edu/coh-metrix/

[24] https://sifnos.iwm-tuebingen.de/ctap/

but also leverages computational tools to reveal instructional needs (Lukwaro et al. (2024)) and potential sociodemographic disparities that current systems often overlook (Hilte et al. (2020); Darkenwald-DeCola (2021); Dowell and Kovanović (2022)).

Within this framework, classical supervised ML experiments are designed to support two aims. First, they seek to approximate the classification logic behind ACCUPLACER. As previously noted, the algorithm used by this system remains proprietary and undisclosed, posing constraints on the transparency and auditability of placement outcomes. Second, ML techniques are applied to a series of classification tasks that lie at the core of this investigation, using both the novel DevEd-inspired annotated features and those automatically extracted via NLP tools. These experiments seek to (i) identify the most informative linguistic features that could enhance classification accuracy (CA) beyond what ACCUPLACER currently achieves, and (ii) determine which ML algorithms yield the best classification performance, all in an effort to support more interpretable and fairer assessment practices in DevEd. The data-mining tool ORANGE[25] (Demšar et al. (2013)) has been selected to support the ML experiments due to its accessible interface and robust suite of well-known ML algorithms.

Beyond NLP tools and ML models, recent advances in the design and development of Large Language Models (LLMs), such as ChatGPT[26] (GPT-3.5-turbo and GPT-4o), Gemini[27] (Pichai and Hassabis (2023)), and Claude[28] (Anthropic (2024)), offer new possibilities for rethinking writing assessment and placement in DevEd. This thesis investigates the use of LLMs, at least as an initial step, toward advancing that broader objective, drawing on their capacity to interpret, explain, and generate language in contextually adaptive ways (Adiguzel et al. (2023)). Recent studies suggest that LLMs can match or even outperform traditional automated systems in text classification and feedback (Xiao et al. (2025)), while offering greater transparency and interpretability through prompt-based integration (Phelps et al. (2024); Wang et al. (2024b)). Higher education institutions are beginning to pilot these tools for writing support and automated feedback (Nikolic et al. (2024)), but further research is needed to assess their reliability, alignment with institutional standards, and responsiveness to equity goals (Porayska-Pomsta and Rajendran (2019)).

This introduction has succinctly outlined pressing issues in language assessment and placement for English (L1) speakers entering DevEd programs, particularly those stemming from the limitations of automated classification systems, in this case ACCUPLACER. Guided by the central question of *how writing proficiency can be more accurately and equitably assessed*, the following section introduces four research questions aligned with the linguistic and computational framework presented here.

[25] https://orangedatamining.com/
[26] https://openai.com/api/
[27] https://deepmind.google/technologies/gemini/
[28] https://docs.anthropic.com/en/home

Research Questions and Thesis Organization

This investigation of language assessment and placement in the context of DevEd, English (L1) writing, is guided by the following four research questions:

RQ1. What lexical, syntactic, and discourse patterns characterize the writing of community college English (L1) speakers placed into DevEd courses?

RQ2. To what extent can linguistically grounded features, both human-annotated and automatically extracted (using NLP tools), improve placement accuracy when modeled using ML?

RQ3. What disparities in placement outcomes emerge across sociodemographic lines such as race and gender, and how can linguistically grounded features improve classification accuracy to promote more equitable DevEd placement?

RQ4. What is the potential of LLMs to serve as complementary tools to existing automated placement systems?

To address these questions, nine studies were conducted. Each study resulted in a double-blind, peer-reviewed paper, published in an international journal, or presented at an international conference. Together, these studies are organized into four interconnected sections, each aligned with one of the research questions and contributing evidence-based insights to support more transparent, accurate, and equitable placement practices.

Section 1, *Building the Linguistic Framework: Annotation, Phraseology, and English Native Language Proficiency*, introduces a linguistic annotation scheme and a phraseological analysis framework tailored to DevEd contexts, offering insights into students' lexical and syntactic preparedness for writing-intensive coursework. The papers included in this section, published between 2022 and 2024, capture the early stages of this investigation, focusing on identifying lexical, syntactic, and discourse markers that indicate writing proficiency and reflect language readiness in native English speakers before beginning an academic program. This section defines the linguistic scope of the thesis and anchors it in theory and annotation methodology, supporting the first research question (RQ1).

Section 2, *Advancing Writing Proficiency Classification with NLP-extracted Features and Machine Learning (ML) Models*, marks a progression from linguistically informed annotation and feature quantification toward the application of supervised ML algorithms for text classification. The work captured in this section scales up the analysis to evaluate the classification accuracy and the performance of different ML models across different linguistic feature sets. In response to the second research question (RQ2), this section investigates how ML can support more precise assessments of writing proficiency than the currently used automated placement

system —ACCUPLACER. It integrates automated linguistic analysis using NLP tools such as COH-METRIX and CTAP, which extract features related to syntactic complexity, lexical sophistication, and cohesion, with a manually annotated set of DevEd-Specific (DES) features. These features are introduced and tested to assess their predictive value, offering a linguistically grounded alternative that aims to improve classification accuracy beyond what ACCUPLACER currently achieves. While some of the papers in this section are more recent than those in Section 3, they appear here because they establish the computational foundation necessary to address broader concerns of accuracy, fairness, and systemic consistency explored later in the thesis.

Section 3, *Evaluating Equity and Classification Consistency in Developmental Education Placement Models*, addresses the third research question (RQ3). It builds on the modeling approach presented in Section 2 by examining the extent to which ACCUPLACER aligns with human assessments of writing proficiency. This alignment (or lack thereof) sheds light on how accurately the system captures the nuanced linguistic patterns that trained human raters are better equipped to identify, and how this affects the reliability of current placement practices in higher education. Additionally, this section investigates how sociodemographic factors, particularly race and gender, intersect with placement outcomes. It pivots to critical questions of fairness and inclusiveness: even if writing proficiency can be modeled with greater precision, are existing tools and processes intentionally designed to reflect both strengths and deficiencies in student writing? And do they overemphasize deficits, especially among historically underrepresented groups? Together, the two studies in this section call for placement models that are not only more accurate, transparent, and consistent, but also more pedagogically sound and responsive to the linguistic needs of diverse learners.

Section 4, *Exploring Next-Generation Writing Placement: Leveraging LLMs for Developmental Education*, addresses the fourth and last research question (RQ4) of this thesis. It extends the contributions of this thesis by introducing Large Language Models (LLMs) into the field of language proficiency assessment and placement. This section investigates whether LLMs, specifically Claude, Gemini, ChatGPT-3.5, and ChatGPT-4o, can offer a more adaptive, transparent, and accurate alternative for automated writing assessment and placement rather than relying solely on commercially developed systems like ACCUPLACER. Beyond classification accuracy, the study in this section briefly notes that LLM-generated feedback is potentially more informative than the high-level, generic comments provided by ACCUPLACER, suggesting opportunities to support literacy instruction and targeted interventions in DevEd. This section is intentionally placed last, representing a forward-looking and exploratory approach that advances the ongoing effort to develop more responsive and inclusive placement practices.

These four sections portray the evolution of this research itself: from building a linguistic framework (Section 1), to modeling and validating it through NLP and ML techniques (Section 2), uncovering implications of current placement systems and processes on accuracy, fairness, and transparency (Section 3), to, finally, exploring innovative placement approaches through

generative Artificial Intelligence (AI) (Section 4). Grouping the papers thematically allows this thesis to tell a cumulative story, aligning the research's development with its linguistic and pedagogical contributions.

Section 1
Building the Linguistic Framework: Annotation, Phraseology, and English Native Language Proficiency

Section 1 introduces a linguistically grounded approach to analyzing student writing through annotation schemes and tagging protocols. It addresses the first research question (RQ1), *What lexical, syntactic, and discourse patterns characterize the writing of community college English (L1) speakers placed into DevEd courses*?, through three published papers.

The first paper, titled *Charting the Linguistic Landscape of Developing Writers: An Annotation Scheme for Enhancing Native Language Proficiency* (Da Corte and Baptista (2024a)), presents a framework, guided by a Corpus Linguistics approach (McEnery et al. (2006); Xiao (2010); Durrant (2022)), designed to capture the linguistic patterns of developing writers. The framework was applied, through a rigorous tagging protocol, to a pilot corpus of 100 essays, all written by native English-speaking students and automatically classified by ACCUPLACER into two levels: DevEd Level 1 or DevEd Level 2. The essays were randomly selected from a broader pool of 290 students who completed the ACCUPLACER test during the 2021–2022 academic year, from Tulsa Community College's (TCC) standardized entrance exam database.

Although modest in size, the corpus of 100 essays, balanced by level, comprised 27,916 tokens. On average, each text had 279 tokens, with the longest text containing 422 tokens and the shortest 95. This corpus offered a valuable foundation for (i) understanding how students placed in DevEd courses express themselves in writing, and for (ii) identifying which students may be most in need of additional academic support (and how best to provide it). To maintain a strictly linguistic focus at this stage, demographic variables such as gender and race were intentionally excluded.

To systematically identify and mark the linguistic features of developing writers, a set of annotation guidelines was created by two expert linguists. The guidelines were developed as part of a pilot annotation study in which three sample essays from the same database (and the same epoch) were manually analyzed. From this pilot, 21 features were identified and grouped into four linguistic categories: orthographic, grammatical, lexical-semantic, and discursive[29]. Some of these features signaled areas of writing deficiency, while others captured more advanced markers of proficiency. Together, they formed the foundation for a multilayered annotation framework that incorporated patterns commonly discussed in the literature (e.g., punctuation errors, parts-of-speech [PoS] omissions), while also introducing novel features not yet studied, such as *fictional we* (fictional representation of a person with the pronoun *we*) and *fictional you* (fictional representation of a person with the pronoun *you*).

[29] All details corresponding to the features included under each pattern and their definitions are included in **Appendix B**

With the guidelines and features in place, the annotation of the corpus was conducted by two trained raters: one male and one female, both native English speakers, graduates in Psychology, and academic advisors at TCC. Notably, both were familiar with DevEd principles and the ACCUPLACER system. To safeguard objectivity, (1) essays were anonymized, (2) raters were not privy to demographic information, and (3) multiple calibration sessions were held to ensure consistency in the annotation process. The raters underwent extensive training, which encompassed (1) an introductory session to present the annotation framework and guidelines; (2) a supervised annotation (practice) session using the three sample essays; and (3) a review session with one of the expert linguists to resolve ambiguities that resulted in the annotation process. These same raters contributed to subsequent studies throughout this investigation, adding methodological consistency to the experiments.

On average, annotators reported spending 17 minutes annotating each essay. A total of 14,135 tags were applied across the corpus. Between the two annotators, there was an overall difference of 1,145 tags. At a more granular level, Annotator 1 applied 65 tags per essay on average, while Annotator 2 applied 76, which is the equivalent of 11 more tags than Annotator 1. Nevertheless, the number of unique tags used per text remained closely aligned, differing by only two. Breaking down the distribution by pattern category, orthographic patterns accounted for the largest share, with 6,265 tags (44.32%). Grammatical patterns followed with 3,183 tags (22.52%), then discursive patterns with 2,713 tags (19.19%), and finally lexical-semantic patterns with 1,915 tags (13.55%). The most frequent tags (presented in all caps) within each category are as follows:

- **Orthographic:** CONTRACT (use of contractions; feature not employed in formal academic writing) with 807 instances, ORT (typography errors) with 2,872 instances, and PUNCT (incorrect or omitted punctuation) with 2,344 instances.
- **Grammatical:** WORDADD (unnecessary word insertion) with 1,060 instances and WORDOMIT (part of speech omitted) with 1,137 instances.
- **Lexical-semantic:** MWE (multiword expressions) with 431 instances and PRECISION (imprecise word use) with 725 instances.
- **Discursive:** FICTIONAL WE with 831 instances and FICTIONAL YOU with 1,403 instances.

To assess tagging consistency, a quality assurance review was conducted to identify areas of disagreement between the annotators and better understand the possible nature and scope of potential tagging inaccuracies. Interrater reliability was then calculated using Krippendorff's Alpha, both at the individual essay level and across the corpus as a whole. The average K-alpha coefficient for individual texts was 0.36, indicating a *fair* level of agreement. When aggregated across all 100 essays, the K-alpha coefficient rose slightly to 0.40, suggesting a *moderate*

level of consensus between annotators. Although modest, the results indicated that annotators understood the guidelines and the linguistic features they were tasked with capturing. They also highlight areas for refinement, particularly in clarifying overlapping definitions and offering more examples, to support greater consistency and reproducibility in future annotations. Additional steps taken to ensure annotation quality assurance are described in the paper supporting this study.

Because writing assessment should capture both strengths and weaknesses in text production, this thesis builds on its initial annotation framework through a second study, titled *A Phraseology Approach in Developmental Education Placement* (Da Corte and Baptista (2022)). This study (i) explores the role of multiword expressions (MWE) in characterizing proficient English (L1) writing discourse, and (ii) signals linguistic markers of emerging proficiency through the identification and tagging of MWE, which have been recognized as indicators of higher-level writing ability (Savary et al. (2019); Kochmar et al. (2020); Hinkel (2022); Esfandiari and Ahmadi (2023)).

In this second study, a custom annotation scheme for MWE was developed by the same two expert linguists who designed the earlier 21-DES feature framework, drawing on both prior research and established linguistic frameworks (Nam and Park (2020); Pasquer et al. (2020); Dahunsi and Ewata (2025)). The scheme drew on dictionary-based resources and classification principles to identify and categorize diverse MWE types within the corpus, ensuring a systematic and linguistically grounded approach. The MWE categories, as presented in (Kochmar et al. (2020)), and the caveats by (Laporte (2018)) were particularly insightful in the development of such a scheme. A total of nine MWE categories[30] were identified: (1) compound adjectives <ADJ>, (2) compound adverbs <ADV>, (3) compound conjunctions <CONJ>, (4) compound nouns <N>, (5) compound prepositions <PREP>, (6) named entities <NE>, (7) phrasal verbs <PHV>, (8) support verb constructions <SVC>, and (9) verbal idioms <VIDIOM>.

After developing the MWE annotation scheme, an expanded corpus of 186 texts was prepared, balanced by ACCUPLACER classification levels (93 from DevEd Level 1 and 93 from DevEd Level 2) and standardized for sample size by including only the first 100 tokens per text (corpus size: 18,600 tokens total). Prior to this stage, additional texts had become available through TCC's entrance exam database and were, therefore, included. This expanded corpus comprised the original 100 texts previously assembled for this investigation, along with 86 newly added samples. The additional texts were randomly selected from the same institutional entrance exam database and time period (2021–2022 academic year), following the same selection protocols approved by TCC's Institutional Review Board (IRB).

The annotation scheme was then manually applied to the full corpus by the two expert linguists previously mentioned, both of whom had extensive familiarity with MWE and MWE types. The annotation process yielded a total of 1,260 MWE, including repeated instances of the

[30]The definitions adopted for each MWE type, with their respective tag, are included in **Appendix B**

same expression across different texts. Upon review of the annotated data, MWE usage appeared more pronounced in texts classified as DevEd Level 2 (by ACCUPLACER). After duplicates were removed, the list of identified MWE contained 600 unique instances, most commonly observed in the following categories: PHV (167 instances); N (158 instances), ADV (110 instances), and VIDIOM (77 instances). Examples of MWE and their categories identified in the corpus included: compound nouns, such as *golden rule*; compound adverbs, such as *by the way*; verbal idioms, like *see the bigger picture*; as well as phrasal verbs, such as *mess up*. Additional examples of MWE (and their types) can be found in **Appendix C**.

The corpus annotated with these MWE-based features was used then to run a series of ML experiments, envisaged as a classification task, aimed at evaluating the potential role of MWEs in predicting writing placement outcomes. While several of the studies already mentioned have proposed frameworks for analyzing the lexical and syntactic features of MWE, it is not known, to the best of one's knowledge, whether automated placement systems, more precisely ACCUPLACER, take these features into account in the placement process (Zachry Rutschow et al. (2021)). Da Corte and Baptista (2022) took on this challenge to investigate whether MWE usage could meaningfully inform DevEd placement outcomes. The data-mining tool ORANGE was chosen for these experiments due to its accessibility and robust ML functionality. Simple and adaptable models[31], such as Naïve Bayes (NB), Neural Networks (NN), and Decision Tree (DT), were applied to assess whether the integration of MWE could improve placement accuracy.

The classification task compared two versions of this corpus: one unannotated, which served as the baseline, and another one with MWE treated as single tokens[32]. To configure the baseline experiment, a basic preprocessing setup was implemented. Tokenization was configured to retain punctuation, recognizing it as a salient linguistic feature that continues to develop in DevEd student writing and may hold predictive value for placement classification. No part-of-speech (PoS) tagger was applied at this stage. In the baseline experiment, the NB was the best-performing model, achieving a classification accuracy (CA) score of 74.2%. Although the CA score is relatively high, it also means that three out of ten students are misclassified in DevEd courses, based on the written texts they produced. When the initial baseline configuration was enhanced by using the same corpus but with its MWE joined as single tokens (and no PoS tagger), the NN algorithm achieved a CA score of 82.3%, representing an 8.1% improvement from the baseline.

These results suggest two important takeaways: (i) MWE influence how writing is represented and interpreted in classification tasks, and (ii) their presence may offer insight into the broader role of formulaic language in student writing, particularly in a DevEd context. While more research with a larger corpus is needed to fully understand the impact of MWE on place-

[31]The definitions and functionalities of the ML models referenced throughout this thesis were obtained from ORANGE's Widget Catalog.

[32]To ensure that the ORANGE data-mining tool recognized each MWE as a single token, its elements were manually joined in the input text with underscores (e.g., *by_the_way*). This step had to be done manually due to the high frequency of typos and spelling errors in the essays.

ment, this study marks an important first step in bringing phraseological patterns into writing assessment frameworks.

While the primary focus of this thesis is on English (L1) writing, the MWE annotation framework initially developed and applied to the English corpus was later adapted to a comparable Spanish-language developmental writing context at the same institution, demonstrating its cross-linguistic relevance and adaptability. This extension, presented in the third paper titled *Multiword expression tagging of Spanish native and non-native speakers' written essays in a grammar and composition developmental course* (Da Corte and Baptista (2023)), was motivated by the recognition that the use of MWE, and the way they signal proficiency, applies across different language backgrounds (Siyanova-Chanturia and Spina (2020)). This cross-linguistic application also aligns with findings from second language acquisition research, which mentions notable distinctions in how native and non-native speakers use MWEs (Pastor (2017)). Because of this, the frequency and accuracy of MWE usage have been linked to broader indicators of language proficiency, and are considered valuable features for automated essay scoring and classification, especially when assessed alongside grammar, morphology, coherence, and sentence structure (Wilkens et al. (2022)).

In this third study, the developmental writing trajectories of native and non-native Spanish speakers enrolled in a Spanish college-level grammar and composition course were examined. A second corpus of Spanish texts was assembled. The corpus consisted of a total of 13,606 tokens, comprising 5,839 tokens from 16 essays written by native speakers and 7,767 tokens from 25 essays written by non-native speakers. It encompassed student essays written at two different points in the semester: one at the outset of the course and the other after five weeks of instruction. The annotation scheme included eight MWE categories: (1) adverbio (ADV) – adverbial expressions; (2) sustantivo común compuesto (N) – compound common nouns; (3) locución conjuntiva (CONJ) – conjunctive phrases; (4) entidad nombrada (NE) – named entities; (5) preposición (PREP) – prepositional phrases; (6) pronombre (PRON) – pronouns; (7) construcción con verbo soporte (SVC) – support verb constructions; and (8) expresión fraseológica verbal (VIDIOM) – verbal idiomatic expressions.

Each essay was manually annotated by the same two expert linguists, from the previous study, who have extensive training in corpus annotation and familiarity with the multiword expression (MWE) taxonomy used in (Da Corte and Baptista (2022)). The total count of (unduplicated) MWE by non-native speakers was 457. The total count of (unduplicated) MWE by native speakers was 322. The total overall MWE count from both groups, native and non-native speakers, was 779, with compound nouns (N; 293 instances), compound adverbs (ADV; 208 instances), and support verb construction (SVC; 113 instances) being the most frequently used and tagged categories of MWE[33].

[33] For illustration purposes, examples of the MWE expressions annotated on both native and non-native spearker's texts are detailed in **Appendix C.1**.

This cross-linguistic extension of the MWE annotation framework aims to evaluate how these features function as indicators of writing proficiency in Spanish, mirroring patterns previously observed in the English corpus. By tracing how specific MWE appear, cluster, or shift between writing samples collected at different points in the semester, this study investigates how lexical complexity and formulaic language reflect developmental gains and how these gains may impact students' proficiency level classification. Based on this aim, three ML experiments[34] were conducted using the same data-mining tool, ORANGE, and similar ML algorithms (e.g., Naïve Bayes [NB]; Neural Network [NN]), as in the prior English-based study. To prepare the annotated texts for the ML experiments, the 779 MWE identified across the corpus were linked with underscores so they would be treated as single tokens in the ORANGE environment. In addition to joining them, full MWE tagging was applied, assigning each expression a type label (e.g., ADV, SVC).

The experiments were as follows: (1) a baseline experiment, Experiment 1, using untagged texts, with no MWE joining or categorization; (2) a second experiment, Experiment 2, with MWE joining and categorization. Additional subexperiments were carried out here, following a one-out strategy: one MWE type was removed at a time, and then combinations of them were excluded. These subexperiments enabled the assessment of both the individual and combined impact of MWE types on classification outcomes; and (3) a third experiment, Experiment 3, removed the MWE type labels (i.e., no explicit categorization), while still preserving the joined format using underscores. This configuration allowed the assessment of the impact that lexical bundling alone had on classification accuracy.

In Experiment 1 (untagged texts with no MWE integration), the NN was the best-performing model among those tested, achieving a classification accuracy (CA) of 79.7%. This model was used as the reference ML algorithm in subsequent experiments, based on its stable performance and comparable findings from a prior study (Da Corte and Baptista (2022)). After establishing this baseline, Experiment 2 introduced MWEs joined by underscores and tagged by type. This integration, in itself, led to a notable improvement in NN's performance, with the CA rising to 82.4%, representing a nearly 3% increase.

Within Experiment 2, additional subexperiments were conducted using a one-out strategy to better understand the impact of each MWE type on model performance. Among the MWE types, verbal idioms (VIDIOM) and compound nouns (N) had the most significant impact on the model's performance. Removing either from the classification task resulted in a noticeable deterioration in the CA score. The removal of each MWE type (VIDIOM and N) resulted in a 5.4% decrease in accuracy, *ex aequo*. On the contrary, when named entities (NE) or pronouns (PRON) were excluded, the NN model's performance improved, suggesting that these features may hinder the classification task. Accuracy rose to 83.8% when NE were removed, the same

[34]For consistency and clarity, all *experiments*, *classification strategies*, and *steps* in this final thesis are identified numerically, e.g., (1), (2), (3), . . .

as when PRON were removed. When VIDIOM and N were removed as a bundle (both features were removed simultaneously), the model's CA also decreased to 78.4%.

While the overall distribution of MWE types did not differ substantially between native speakers and non-native speakers, confirmed by a Pearson correlation score of 0.894, notable distinctions emerged in the use of VIDIOM and N. These differences, though observed within a relatively small corpus, suggest that specific MWE categories, particularly those reflecting idiomatic and nominal phraseology, may serve as stronger indicators of near-native writing proficiency.

Lastly, in Experiment 3, MWE remained joined and were treated as single tokens; however, all eight categorical (type) tags were removed. Thus, while the model could still recognize the presence of an MWE, it did not know its specific type. The results of this experiment revealed a slight decrease in the NN's accuracy performance, 1.3% lower (to 78.4%) than the baseline (79.7%), and 4% lower than Experiment 2 (82.4%). These findings suggest that retaining MWE-type information likely supports more accurate classification, though, as already mentioned, further research with a larger corpus is needed.

References to the papers in order of presentation:

Paper 1:
Da Corte, M. and Baptista, J. (2024). **Charting the Linguistic Landscape of Developing Writers: An Annotation Scheme for Enhancing Native Language Proficiency.** In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3046–3056, Torino, Italia. ELRA and ICCL.

Paper 2:
Da Corte, M. and Baptista, J. (2022). **A Phraseology Approach in Developmental Education Placement.** In Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain, pages 79–86.

Paper 3:
Da Corte, M. and Baptista, J. (2023). **Multiword Expression Tagging of Spanish Native and Non-Native Speakers' Written Essays in a Grammar and Composition Developmental Course**. Romanica Olomucensia, 35(1):23–40.

Charting the Linguistic Landscape of Developing Writers: an Annotation Scheme for Enhancing Native Language Proficiency

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

This study describes a pilot annotation task designed to capture orthographic, grammatical, lexical, semantic, and discursive patterns exhibited by college native English speakers participating in developmental education (DevEd) courses. The paper introduces an annotation scheme developed by two linguists aiming at pinpointing linguistic challenges that hinder effective written communication. The scheme builds upon patterns supported by the literature, which are known as predictors of student placement in DevEd courses and English proficiency levels. Other novel, multilayered, linguistic aspects that the literature has not yet explored are also presented. The scheme and its primary categories are succinctly presented and justified. Two trained annotators used this scheme to annotate a sample of 103 text units (3 during the training phase and 100 during the annotation task proper). Texts were randomly selected from a population of 290 community college intending students. An in-depth quality assurance inspection was conducted to assess tagging consistency between annotators and to discern (and address) annotation inaccuracies. Krippendorff's Alpha (K-alpha) interrater reliability coefficients were calculated, revealing a K-alpha score of k=0.40, which corresponds to a moderate level of agreement, deemed adequate for the complexity and length of the annotation task.

Keywords: annotation scheme, developmental education, native language proficiency

1. Introduction and Objectives

Developmental courses or college-ready models of education provide supplemental instruction in key foundational areas, such as English writing, to students performing below academic standards before participating in college-bearing courses (Mazzariello et al., 2018). Placement in DevEd is based on the automatic assessment and scoring of linguistic features extracted from standardized written assignments, administered as an entrance exam through ACCUPLACER[1].

The entrance exam requires students to compose a short essay, typically between 300 to 600 words, addressing a specific writing prompt, e.g., *Success in Life*; *Making Mistakes*. Upon submission of the essay, ACCUPLACER automatically assesses and assigns it a classification level: DevEd Level 1, 2, or College level (demonstrating no need for DevEd). Currently, 46% of students who complete this assessment are placed in DevEd, based on statistics (from the Fall 2022 semester) reported by Tulsa Community College (TCC)[2], the higher education setting where this study took place.

The literature is limited on how ACCUPLACER performs this complex task and, most importantly, what linguistic features are included in the classification process and how the classification constructs are devised to automatically place students in the respective courses (Perin et al., 2015). The challenges associated with these limitations raise questions of how accurate the placement of these students is and if they are receiving adequate support to aptly communicate and participate in an academic program (Nazzal et al., 2020).

Following an annotation scheme for enhancing language proficiency, it is expected to more systematically outline (and categorize) the linguistic features of the texts produced by this population undergoing ACCUPLACER assessment and placement. Additionally, it is expected to explore other salient features that the literature has not fully addressed and investigate if they are indicative of students' writing proficiency levels. The ultimate goal is to (i) contribute towards the establishment of a reference corpus of annotated essays to serve as learning material for ACCUPLACER's machine/deep learning, (ii) produce a more accurate estimation of the classification and course placement of students, and (iii) further align students' college-readiness skills with the literacy demands of higher education.

2. Related Work

The placement of students in DevEd courses in the United States has been a topic of debate, primarily centered around the validity of test results (Perin et al., 2015; Griffiths II, 2019). Nazzal et al. (2020) explains some of the limitations of commonly used standardized placement exams, such as Ac-

[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/

20
LREC-COLING 2024, pages 3046–3056

CUPLACER. These limitations include (i) the structure of these exams, often consisting of multiple-choice questions and a few essay prompts, and (ii) how these systems are trained to automatically place students based on the narrowed conceptualization of the writing process they portray (Hughes and Li, 2019).

The question of whether standardized placement exams are the best way to accurately measure the need for students to gain or remediate basic academic skills continues to be discussed among community college researchers (Cullinan and Biedzio, 2021; Klausman and Lynch, 2022). It is estimated that these exams misplace about 30% to 50% of students (Hassel and Giordano, 2015) and that the scores they produce do not correlate with students' success in college (Hassel and Giordano, 2015).

Because of these limitations, a more precise depiction of the lexical patterns exhibited by students with English as their L1 is needed. Nazzal et al. (2020) focused on identifying writing strengths and weaknesses of community college students and relied on a text-based *academic writing assessment* (AWA) that prompted participants to read, interpret, and synthesize complex texts to build an argument and present it in written form. Results from the study confirmed that, with non-standardized writing assessments created at the institutional level, identifying groups with varying writing proficiency levels becomes more effective and provides more complete and "nuanced information" on writing abilities.

Graesser et al. (2004); Perin and Lauterbach (2018) used Coh-Metrix (McNamara et al., 2006)[3], a natural language processing (NLP) automated scoring engine, to assess the written work of low-skilled students after completing specific writing assignments. The assessment considered linguistic variables, including connectives, lexical overlap, logical operators, anaphoric reference, and syntactic complexity. The motivation for using Coh-Metrix to detect (poor) proficiency levels was substantiated. The researchers suggested further investigations "to verify the accuracy of human scores and the theoretical relevance of automated linguistic scores" (Perin and Lauterbach, 2018, p.73) to inform instructional practices in the classroom.

Chen and Meurers (2016) developed a web-based tool, Common Text Analysis Platform (CTAP), using NLP tools to enhance the linguistic features identified by human annotators from written texts. Many of the criticisms the authors present regarding the accuracy, adaptability, reliability, and validity of results relate to others already advanced by the literature regarding classification systems such as ACCUPLACER.

Abba (2015); Abba et al. (2018) also used Coh-Metrix to explore lexical patterns among different student groups (L1, L2, and Generation 1.5; the latter are students educated in the U.S., but who do not have English as L1) and compared them to lexical and syntactic features of proficient writing. The comparison focused on noun overlap in adjacent (and all) sentences, argument overlap, lexical diversity, pronoun incidence, length of sequential word strings, familiarity with content words, modifiers in a noun phrase, as well as noun and verb phrase density. The study concluded that word familiarity and word polysemy suggested the most linguistic differences among the groups.

Staples and Reppen (2016); Duran (2017); Khan (2019); Kyle (2019); Gilquin and Granger (2022) relied on annotated corpora to investigate differences and variations of lexico-grammatical choices made by developing writers when completing an array of writing tasks. Gilquin and Granger (2022), in particular, indicated that annotated corpora provide learners with the opportunity to compare their writing with that of expert (or native-level) writers or consult with "learner corpus where errors have been annotated, [...] to correct their own interlanguage features (misuse, overuse, underuse) and thus improve their writing." [p.2].

These studies offer a reference framework for this study, emphasizing how the intricacies of the English language impact the assessment of students' written communication skills and their course placement prior to beginning their academic career.

3. Methodology

3.1. Corpus

A sample of 103 text units (essays), written in English, was randomly selected from a population of 290 community college intending students during the 2021-2022 academic year. These essays were produced without a time limit or editing tools at TCC's proctored testing center.

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. Based on the existing literature, this automated system does not offer specific definitions for the levels, as these are customized by individual institutions. For this study, the levels were defined as follows: *Level 1* entails a text unit showing a need for development in the general use of English, including grammar, spelling, punctuation, and the structure of sentences and paragraphs. *Level 2* suggests a text unit requiring targeted support in particular aspects of the English language, such as sentence structure, punctuation,

[3] http://tool.cohmetrix.com/

editing, and revising. Other metadata, including demographics (gender, race, among others), was ignored at this stage. Table 1 shows the corpus size in token numbers.

Parameter	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: Size of the corpus for pilot annotation task.

Of the 103 sample units, 3 were used to train the annotators, while 100 were used for the core annotation task, equally divided by level (50 for Level 1 and 50 for Level 2). These 100 sample units constitute the initial seed for a corpus capturing the language variety of community college students and underpin the annotation task.

3.2. Annotation Scheme

Two linguists with extensive experience in identifying and categorizing linguistic features in written corpora designed an annotation scheme[4] to standardize and maintain consistency throughout the annotation process (Da Corte and Baptista, 2024). Some of the linguistic criteria referenced are supported, among others, by McNamara et al. (2006); Martinez (2013); Omidian et al. (2017); Thomson (2017); Da Corte and Baptista (2022); Glass (2022), and Senaldi et al. (2022); while others, such as the *Fictional You* (imaginary representation of a person by the pronoun *you*.) and *Fictional We* (similar to the previous feature but using the pronoun *we*.), emerged through direct inspection of the corpus as potential indicators of developing writers' proficiency levels. These latter features are unique in that they have not been fully explored in the available literature.

Another distinct aspect of this scheme was considering and including multiword expressions (MWE) as proficiency-signaling features. Given that numerous MWE frequently exhibit meanings not directly inferred from their composition. Since they also account for a substantial share of textual elements across various texts, they are anticipated to significantly influence the representation of data within texts. Consequently, this influence is likely to enhance tasks involving text classification (Da Corte and Baptista, 2022).

The annotation scheme includes an exhaustive list of 28 features across 4 textual pattern categories, with various examples, and, additionally, 4 broader textual criteria to holistically assess the sample units upon the conclusion of the annotation process. The 4 textual patterns included in these guidelines are classified into 2 main types: (i) patterns that signal errors and indicate a deviation from proficiency standards and (ii) patterns that signal proficiency. The term *proficiency standards* is here used to refer to accepted norms or expectations for the use of the English language.

The patterns deviating from proficiency standards, outlined in ascending order based on their degree of complexity, are:

(1) **Orthographic patterns:** patterns representing the foundational language skills needed to represent words and phrases, which, in turn, supports learning tasks associated with vocabulary and grammatical knowledge (Kim et al., 2017). These patterns (with their respective tag) consist of 8 linguistic features: *grapheme addition* <ADD>; *grapheme omission* <OMIT>; *grapheme transposition* <TRANSPOSE>; *grapheme capitalization* <CAPS>; *word split* <WORDSPLIT>; *word boundary merged* <WORDMERGED>; *punctuation used* <PUNCT>; *contractions* <CONTRACT>. For this annotation task, all features at the grapheme level (4) were tagged as <ORT>. A subsequent analysis will look into these features at a more granular level.

(2) **Grammatical patterns:** patterns evidencing the quality of a writer's text production and have been previously used in the automatic prediction of language proficiency levels through systems such as ACCUPLACER. The literature deems these patterns as valuable features used in the analysis of learner corpora, which can then be used as a "standard resource for empirical approaches to grammatical error correction," (Glass, 2022; Dahlmeier et al., 2013, p.22). These patterns (with their respective tag) consist of 9 linguistic features: *word omitted* <WORDOMIT>; *word added* <WORDADD>; *word repetition* <WORDREPT>; *verb tense* <VTENSE>; *verb disagreement* <VDISAGREE>; *verb mode* <VMODE>; *verb form* <VFORM>; *adjective-adverb interchange* <INTERCHANGE>; *pronoun-alternation referential* <ALTERN>.

(3) **Lexical & semantic patterns:** patterns contributing to the structuring and formation of statements that, when linked or combined with other statements, shape a writer's discourse. These patterns could serve as indicators of how writing skills develop over time. These patterns (with their respective tag) consist of 4 linguistic features: *slang* <SLANG>; *word precision* <PRECISION>; *mischosen preposition* <PREP>; *connectives* <CONNECT>.

(4) **Discursive patterns:** patterns exhibiting the production of multiple utterances or writer's ability to produce extended text to discuss a topic, reformulate it, support an opinion, and hypoth-

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

esize, among other higher-order thinking tasks. Discursive patterns evidence a writer's ability to write at the academic level or if the development of writing skills is needed. These patterns (with their respective tag) consist of 7 linguistic features: *emphatic do* <EMPH_DO>; *mimesis* <MIMESIS>; *fictional we* <FICTIONAL_WE>; *fictional you* <FICTIONAL_YOU>; *argumentation with reason* <REASON>; *argumentation with examples* <EXAMPLE>; *allegory* <ALLEGORY>.

The structured list of the categories and features proposed is not a closed nor a definitive set and is subject to refinement, which is the goal of this annotation task. Adjustments will be made based on usability and automatic identification through natural language processing (NLP) tools. The same caveat applies to the features included under the patterns signaling proficiency described below.

The patterns signaling proficiency fall under the lexical and semantic category (the definition is the same as the one already mentioned) and include several subcategories of MWE. The caveats by Laporte (2018) and the categories devised by Kochmar et al. (2020) were particularly insightful. No information on the ACCUPLACER strategy for signaling and factoring MWE into the assessment process has been found. The impact of MWE on DevEd placement was investigated in previous studies, such as Nam and Park (2020); Da Corte and Baptista (2022). For this annotation task, all MWE categories (8) were tagged as <MWE>. A subsequent analysis will look into these 8 categories more granularly.

The 28 linguistic features described thus far are designed to be directly tagged on each sample unit during annotation. However, beyond these specific features, there are also 4 broader textual criteria:

(1) **Sentence variety and style** <SENTENCEVAR>: encompasses the repetition (or not) of sentences and strings of sentences and the value that it adds to the discourse.

(2) **Paragraph composition** <PARAGRAPH>: encompasses how structured and cohesive the text is and if it showcases the basic essay outline of a clear introduction, supportive statements, and a conclusion.

(3) **Prompt adherence** <ADHERE>: encompasses how the statements used to support the ideas connect (or not) to the main prompt and/or thesis statement(s) of the text.

(4) **Prompt resumption** <RESUMPTION>: encompasses the development of a topic and/or thesis using supportive statements or evidence.

These criteria enable annotators to holistically evaluate a writer's academic writing ability after annotating each sample text unit and offer a more transparent approach to assessment, integrating elements utilized by ACCUPLACER for placement. Within each textual criterion, a 4-point Likert scale was adopted: 0 - *deficient*; 1 - *below average*; 2 - *above average*; 3 - *outstanding*.

3.3. Annotators and Training

A call for volunteers to participate in the annotation task was disseminated at TCC, with 2 out of 6 respondents selected based on their background, skills, and experience. The 2 annotators, 1 male and 1 female, were (i) native English speakers, (ii) graduates of U.S. higher education in Psychology, (iii) academic advisors at TCC, and (iv) familiar with DevEd principles and the ACCUPLACER system and placement guidelines.

Both annotators previously engaged in another annotation task associated with this study. For the current task, they completed a comprehensive three-day training provided by one of the annotation scheme authors. This training encompassed (i) expectations and ethical considerations, (ii) annotation task procedures, and (iii) detailed instruction on the application of the scheme referenced in Section 3.2, with abundant examples covering all proposed cases under the orthographic, grammatical, lexical, semantic, and discursive already mentioned. This ensured that the annotators were well-prepared to identify all cases.

The training involved two annotation practice rounds on 3 sample texts. The first round was *guided*, while the second was *independent*. After the independent round, a debrief session was organized. In this session, the annotators and the trainer reviewed selected annotations and discussed (and addressed) any discrepancies. Additionally, a timeline was discussed, including a 30-day deadline for completing the annotations and a check-in on day 15. Although the training was voluntary, a modest stipend was given after annotating all 100 sample units. The formal annotation began promptly after the debrief.

4. Annotation Task Assessment

The annotation task was completed as scheduled. Annotators reported spending an average of 17 minutes per essay, which included reading each text sample, identifying and tagging the linguistic features using a simple text editor, and assessing each using the 4 broader textual criteria already discussed in Section 3.2. The time spent per essay is higher than the one reported during the training phase (15 minutes) due to the size and complexity of the samples. Table 2 presents a breakdown of the total tags applied per text per annotator, pointing out some differences and similarities.

A total of 14,135 tags were applied by both annotators across the corpus of 100 texts containing

Parameter	A1	A2
Total tags	6,495	7,640
Average tags/text	65	76
Average unique tags/text	12	14
Minimum total tags/text	15	24
Maximum tags/text	178	196

Table 2: Tags per text per Annotator (A).

27,916 tokens. The tag distribution indicates an overall difference of 1,145 tags between Annotators 1 and 2. At a more granular level, Annotator 2 used 76 tags, on average, 11 more tags per text than Annotator 1 (65). The average count of unique tags used per text by both annotators was quite similar, differing by just 2. The difference indicates that, on average, Annotator 2 identified two additional linguistic features than Annotator 1. The differences in annotation productivity could be attributed to the application of the guidelines rather than a shortfall in training.

The maximum count of tags applied by Annotator 1 in a single text was 178 (for a 341-token text), while for Annotator 2, it was 196 (for a 329-token text). The texts with the least tags had 15 (150-token text by Annotator 1) and 24 (167-token text by Annotator 2), respectively.

4.1. Tagged Feature Distribution

Table 3 lists the tags classified based on the annotation scheme introduced in Section 3.2 and highlights in bold the ones most frequently used by both annotators, with their respective counts, ratios, and sums italicized.

Of the total tags applied, 6,265 correspond to orthographic patterns, representing an average ratio of 44.32% (pattern tags/overall tags applied). Following are 3,183 tags signaling grammatical patterns (22.52%), and 2,713 tags (19.19%) signaling discursive patterns, leaving approximately 1,915 tags (13.55%) for the lexical & semantic patterns category. A small ratio percentage of 0.42%, or the equivalent of 59 tags, was used to signal a re-occurring feature <DUMMY>, not initially defined in the scheme, warranting further exploration. The vast majority of the tags applied, with an average ratio of 96.95%, signaled patterns deviating from proficiency standards, while the remaining 3.05% signaled proficiency.

For many tags, either both annotators have inserted a similar number of annotations, or there is a slightly higher percentage for Annotator 2. The most asymmetric cases in this distribution involve the features in the discursive patterns, in which Annotator 1 identified *emphatic_do* and *mimesis*, but Annotator 2 did not, and vice versa for *allegory*. This minimal discrepancy requires further inspection as it may impact the calculation of interrater reliability coefficients. A total of 9 tags were most frequently used by both annotators across the 4 main pattern categories, as shown in Table 4, with each major pattern represented by at least two tags.

Calculations of the Pearson coefficient (r) (Cohen, 1988) were performed to assess how diverse the use of a tag set was by each annotator per text (for all 100 sample units). A high coefficient of r=0.834 between the two sets of annotations was obtained for the correlation between the ratios of unique tags/total tags per text. It indicates a strong positive linear relationship between the two arrays of ratios. This r value suggests that the overall number of tags applied in relation to the unique tags identified (per text) was generally comparable and consistent between the two annotators. Figure 1 evidences the correlation between Annotators 1 and 2 ratios.

The standard deviation (σ) between the two arrays of ratios was also calculated to provide further insight into this relationship. A value of (σ)= 1.890 was obtained. This value represents the average amount by which individual ratios deviate from the mean of the ratios. A standard deviation of 1.890, for average values of 5.319 (Annotator 1) and 5.3201 (Annotator 2), indicates some variability in how each annotator applied tags.

Together, the correlation coefficient and standard deviation offer a comprehensive view of the annotators' relationship and the consistency of their application of the tag set throughout the corpus, highlighting a *strong but not perfect* alignment.

4.2. Annotation Quality Assurance

An in-depth inspection of tagging consistency between annotators was undertaken to identify potential areas of disagreement and understand the nature and scope of inaccuracies. Such a step is pivotal in deciding which errors were due in part to the annotators, the guidelines, or the task itself and, consequently, better preparing the annotation of the future corpus.

To accomplish this objective, a 10% sample from each tag and annotator was selected from a total of 14,076 tags applied, which will be investigated at a later time, as detailed in Table 3 (This total does not include the 59 <DUMMY> tags applied, which will be investigated at a later time). All smaller sets of tags comprising fewer than 20 instances were manually and systematically inspected in full. It is worth noting that, as elaborated in Section 3.1, the sample text units had already been randomly ordered to be assigned their respective unique identifier. Thus, by design, this subsample is already randomized.

The inspection of this subsample was anchored around the annotation scheme explicitly devised

Patterns	Tag	A1	Ratio	A2	Ratio	Count	Avg.Ratio
	‹**CONTRACT**›	*417*	*6.42%*	*390*	*5.10%*	*807*	*5.71%*
	‹**ORT**›	*1,504*	*23.16%*	*1,368*	17.91%	*2,872*	*20.32%*
Orthographic	‹**PUNCT**›	*1,000*	*15.40%*	*1,344*	*17.59%*	*2,344*	*16.58%*
	‹WORDMERGED›	46	0.71%	43	0.56%	89	0.63%
	‹WORDSPLIT›	75	1.15%	78	1.02%	153	1.08%
	‹ALTERN›	13	0.20%	168	2.20%	181	1.28%
	‹INTERCHANGE›	2	0.03%	6	0.08%	8	0.06%
	‹VDISAGREE›	92	1.42%	95	1.24%	187	1.32%
	‹VFORM›	24	0.37%	50	0.65%	74	0.52%
Grammatical	‹VMODE›	1	0.02%	18	0.24%	19	0.13%
	‹VTENSE›	90	1.39%	238	3.12%	328	2.32%
	‹**WORDADD**›	*389*	*5.99%*	*671*	*8.78%*	*1,060*	*7.50%*
	‹**WORDOMIT**›	*440*	*6.77%*	*697*	*9.12%*	*1,137*	*8.04%*
	‹WORDREPT›	43	0.66%	146	1.91%	189	1.34%
	‹CONNECT›	100	1.54%	215	2.81%	315	2.23%
	‹**MWE**›	*150*	*2.31%*	*281*	*3.68%*	*431*	*3.05%*
Lexical & Semantic	‹**PRECISION**›	*344*	*5.30%*	*381*	*4.99%*	*725*	*5.13%*
	‹PREP›	84	1.29%	217	2.84%	301	2.13%
	‹SLANG›	72	1.11%	71	0.93%	143	1.01%
	‹ALLEGORY›	0*	0.00%	2	0.03%	2	0.01%
	‹EMPH_DO›	2	0.03%	0*	0.00%	2	0.01%
	‹EXAMPLE›	32	0.49%	48	0.63%	80	0.57%
Discursive	‹**FICTIONAL_WE**›	*541*	*8.33%*	*290*	*3.80%*	*831*	*5.88%*
	‹**FICTIONAL_YOU**›	*768*	*11.82%*	*635*	*8.31%*	*1,403*	*9.93%*
	‹MIMESIS›	3	0.05%	0*	0.00%	3	0.02%
	‹REASON›	229	3.53%	163	2.13%	392	2.77%
Other	‹DUMMY›	34	0.52%	25	0.33%	59	0.42%
Total		6,495	100%	7,640	100%	14,135	100%

Table 3: Total tags used per Annotator (A) in the annotation task.

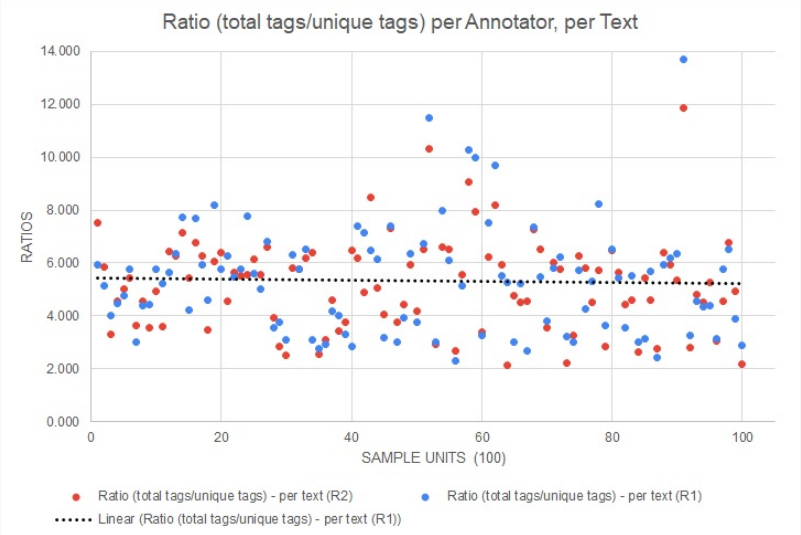


Figure 1: Ratio total tags/unique tags per annotators.

for this task. Of the 664 tags examined for Annotator 1, 43 were incorrectly applied, translating to an error ratio of 6.476%. In contrast, of the 787 tags reviewed for Annotator 2, 104 were incorrect, yielding a 13.214% error ratio. Thus, the error ratio of Annotator 2 was approximately double that of Annotator 1.

This discrepancy prompts two questions:

1. Were there specific features (and their tags) consistently misconstrued?
2. Were any tags (or patterns) notably error-prone?

To address these questions, features with an er-

Patterns	Tag	Total
Orthographic	<CONTRACT>	807
	<ORT>	2,872
	<PUNCT>	2,344
Grammatical	<WORDADD>	1,060
	<WORDOMIT>	1,137
Lexical & Semantic	<MWE>	431
	<PRECISION>	725
Discursive	<FICTIONAL_WE>	831
	<FICTIONAL_YOU>	1,403

Table 4: Most frequently used tags by both annotators.

ror rate exceeding 5% of each tag's total instances were analyzed. Other discrepancies (suggesting a partial mismatch with the guidelines) were also observed during this examination. Again, the primary goal was to explore the consensus between annotators in the annotation process and the usefulness/thoroughness of the guidelines. The absence of a gold standard here is by design. This standard is being produced as the result of this annotation task. A total of 56 **additional errors** were identified, and they have been systematically categorized and coded as follows:

(1) **Incomplete tag** (*m*): a tag was correctly selected but applied incompletely (missing the opening or closing tag or omitting some of its elements like brackets or colons) or with extra characters, e.g., **corpus:**<PUNCT>okay<,okay,>/PUNC>

correct:<PUNCT:okay>,okay,</PUNCT>

(2) **Misplaced feature** (*n*): a tag was correctly selected, but the signaled feature was misplaced, typically suggesting a correction when not required or, on the contrary, not suggesting a correction where there should be one, e.g., (note the example with <WORDADD>) **corpus:** [...] *I've* <WORDADD:have></WORDADD> *know a girl my age* [...]

correct: [...] *I've* <WORDADD>have</WORDADD> *know a girl my age* [...]

(3) **Missing nested tag** (*o*): an identified feature warranted an additional correction, implying a missing nested tag, e.g., **corpus:** [...] *But I think yes fitting in can be valuable in some ways as far as having somone around when* <FICTIONAL_YOU>your</FICTIONAL_YOU> *down* [...] **correct:**[...] *But I think yes fitting in can be valuable in some ways as far as having somone around when* <FICTIONAL_YOU><PRECISION:you are>your</PRECISION></FICTIONAL_YOU> *down* [...]

(4) **Inadequate correction** (*p*): a tag was appropriately chosen, but the provided correction is inadequate, e.g., **corpus:** [...] *After high school I went to platt college and got my medical* <ORT:assitant>assisstant</ORT>*license* [...]

The suggested correction is not spelled correctly, missing the grapheme <s> required in *assistant*.

correct: [...] *After high school I went to platt college and got my medical* <ORT:assistant>assisstant</ORT>*license* [...]

Out of these 56 additional errors, Annotator 1 accounted for 14, while Annotator 2 was responsible for 42, resulting in a 1:3 error ratio between them (with an error rate of 10.13%). Notably, the most common error category for both annotators was missing nested tag (*o*). Since nesting entails a layered hierarchy and interdependency between tags, it becomes a more error-prone process, increasing the likelihood of overlooking or misplacing tags and confirming the difficulties associated with multilayered annotation.

Regarding corpus annotation, it is difficult to claim that a corpus is entirely exempt from errors due to various influencing factors. Nonetheless, the rigorous (and systematic) manual review of the annotations here ensures this task's quality. While no universal standard for error rates exists, as this depends on several factors such as the cognitive load of the task, its length, and the size of the tag set, among others, an error rate below 10% can be deemed acceptable. This error rate is particularly relevant when using annotated corpora in Machine-learning experiments.

The results in this quality assurance step suggest that the annotators had a good understanding of the guidelines provided. Error analysis confirms the need for minor adjustments to reduce feature/tag overlaps, enhancing guideline clarity and annotator training. Next is the calculation and evaluation of the interrater reliability coefficients.

To determine the level of agreement between the two annotators when leveraging the scheme designed for this annotation task, interrater reliability coefficients (IRC) were calculated in two steps using (i) the 28 linguistic features directly tagged on each sample text unit and (ii) the 4 broader textual criteria employed in the holistic assessment of each text.

For the computations of the IRC, the ReCal-OIR tool (Freelon, 2013)[5] was used since it provides the calculation of Krippendorff's Alpha (K-alpha) for *nominal* data for two annotators. An important preprocessing step involved using Python code to tokenize, verticalize, and align each text (100 total). Each annotation (e.g. <TAG>...</TAG>) was considered a single token. Naturally, words were also tokenized, but punctuation signs were kept next to the preceding word (if not a part of a tagged sequence). Through this alignment process, a total of 45,071 tokens were obtained, averaging approximately 451 tokens per text unit. Out of the 45,071 tokens, there was an agreement between Annotators 1 and 2 on 31,638 tokens. This left 13,433 tokens where discrepancies arose, resulting in an observed proportion of agreement of 70.196%.

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

To assess interrater reliability, the calculations were performed individually per text and across all text units. The average K-alpha for individual texts stood at 0.36, indicating a *fair* level of agreement. However, when aggregating the data for a comprehensive analysis of the 100 text units, the K-alpha coefficient rose slightly to 0.40, suggesting a *moderate* level of consensus between the two annotators. The standard deviation (σ) of the text units' K-alpha scores was also calculated, and a σ=0.032 was obtained, indicating *minimal* variability among these scores. Figure 2 presents the individual K-alpha score for each text along with the average of 0.36 (denoted with a red line).

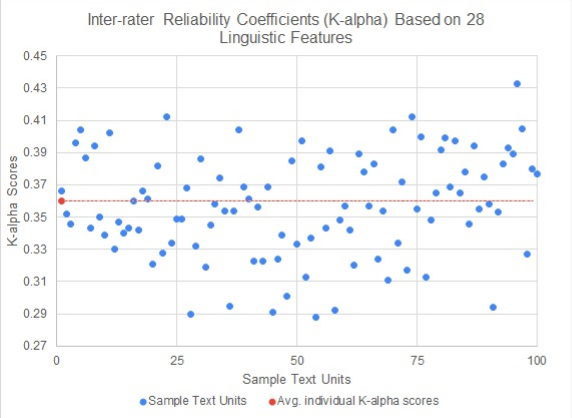


Figure 2: K-alpha scores' distribution around average value.

K-alpha method for *ordinal* data served as the method for reliability calculations for the 4 textual criteria utilized for a holistic assessment of each text unit. The results of this analysis can be found in Table 5.

Criterion	K-alpha
<SENTENCE>	**0.237**
<PARAGRAPH>	**0.295**
<ADHERE>	0.186
<RESUMPT>	0.085
All Criteria	0.244

Table 5: IRC: K-alpha scores for ordinal data.

Results indicate that the two annotators achieved a *fair* interrater agreement for the <SENTENCE> and <PARAGRAPH> criteria; while for the <ADHERE> and <RESUMPT> the values indicate *slight* agreement. Considering all four criteria, the overall K-score of k = 0.235 is interpreted as *fair*. These different results per criteria suggest that while <SENTENCE> and <PARAGRAPH> can be construed as more objective, evidence-based criteria, prompting a more consistent classification from the annotators, <ADHERE> and <RESUMPT> imply assessing the consistency and coherence in the development of the meaning of the prompt topic throughout the essay. This is a cognitively heavier task, consequently resulting in a *suboptimal* agreement between the annotators.

Since the 4-point Likert scale adopted for these (4) criteria (ordinal) may be a too complex task and allows for a greater dispersion of rates, a data transformation was performed, converting the rates into binary values (0 for *negative* assessment, corresponding to *deficient* (0) or *below average* (1) ratings; and 1 to *positive* assessment, corresponding to above *average* (2) and *outstanding* (3) ratings. This data transformation intends to capture an *adjacent* classification agreement and provide a more insightful perception of the data. Results are shown in Table 6.

Criterion	K-alpha
<SENTENCE>	0.127
<PARAGRAPH>	**0.587**
<ADHERE>	-0.005
<RESUMPT>	0.025
All Criteria	0.194

Table 6: IRC with Data Transformation: Binary Values.

When considering all 4 criteria, the overall K-alpha of k = 0.194 represents a decrease from the previous score (k = 0.244) prior to adopting a binary scale. This new K-score evidences a *slight* agreement between the annotators. Within each criterion, 3 out of 4 (<SENTENCE>; <ADHERE>; <RESUMPT>) also evidence a decrease in their respective K-scores, shifting the level of agreement from *fair* to *slight*. On the contrary, <PARAGRAPH> yielded a K-score of k = 0.587, indicating a *moderate* level of agreement. The result of this particular criterion confirms the need to more precisely define and exemplify such constructs to achieve a more consistent and robust classification. This, in turn, is a key step in fleshing out more complex (and less-than-reliable/reproducible) criteria purported by the placement Machine-learning approach used by systems such as ACCUPLACER.

5. Conclusions and Future Work

This study introduced an annotation scheme designed to capture orthographic, grammatical, lexical, semantic, and discursive patterns of college-intending students participating in a DevEd model. These discerned patterns hold potential as indicators for written language proficiency and could optimize placement in DevEd courses, currently performed by automated systems such as ACCUPLACER. Moreover, they pave the way for address-

ing linguistic barriers hindering effective college-level communication.

The annotation scheme was carefully vetted, and rigorous protocols were followed to protect the participants of this study and collected text sample units. The competency of our annotators, as verified by Pearson correlation calculations, demonstrated consistent tagging across all 100 text units. Using K-alpha for nominal data, IRC highlighted a *moderate* level of consensus between the two annotators. In the quality assurance step, an error rate of 10.13% was obtained. This error rate, deemed acceptable based on the complexity of the task, confirms that the annotators had a good understanding of the guidelines provided and suggests minor refinements or revisions to minimize feature/tag overlaps.

The findings of this study led to trimming certain linguistic features registering low incidences, such as <INTERCHANGE>; <VMODE>; <ALLEGORY>; <EMPH_DO>; and <MIMESIS>. Features more susceptible to errors like - <ALTERN>; <VFORM>; <VTENSE>; <WORDADD>; <PRECISION> - will undergo definition refinements, accompanied by richer illustrative examples.

Regarding the 4 broader textual criteria utilized for a holistic assessment of each text unit, K-alpha scores verify that <SENTENCE> and <PARAGRAPH> can be construed as more objective criteria, which can contribute to improving less-than-reliable/reproducible criteria purported by ACCUPLACER.

Future work will focus on expanding the corpus to a size suitable for Machine-learning applications, incorporating relevant features for the task. This expansion will be complemented by the application of natural language processing techniques within a Machine-learning framework, utilizing a data mining toolkit to (i) determine which features are more predictive of the level of English writing proficiency of developing writers and (ii) more accurately determine students' placement in one of the courses' level (Level 1 or 2).

6. Ethical Considerations

This study utilized a systematic sampling method, adhering to TCC's Institutional Review Board (IRB)[6] protocols, 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.

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

7. 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, Rebecca Aberle and Jonah Townsley, who participated in this task, and Chuck Stimac, Application Developer, whose expertise made the systematic analysis of the linguistic features presented in this paper possible. Their thorough work and detailed strategies have significantly contributed to the progress of our research.

8. Bibliographical References

Katherine A Abba, Shuai Steven Zhang, and R Joshi. 2018. Community college writers' metaknowledge of effective writing. *Journal of Writing Research*, 10(1):85–105.

Katherine Anne Abba. 2015. *Community college students' writing: Lexical, syntactic, and cohesion differences in L1, L2, and Generation 1.5 students and examining knowledge of the writing process*. Ph.D. thesis, Texas A&M University, Graduate and Professional Studies.

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.

Dan Cullinan and Dorota Biedzio. 2021. Increasing Gatekeeper Course Completion. Technical report, CCRC, Teachers College, Columbia University.

Miguel Da Corte and Jorge Baptista. 2022. A phraseology approach in developmental education placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86.

Miguel Da Corte and Jorge Baptista. 2024. Annotation scheme for charting the linguistic landscape of developing writers. GitLab repository.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus

of learner English: The NUS corpus of Learner English. In *Proceedings of the 8th Workshop on Innovative use of NLP for Building Educational Applications*, pages 22–31.

Derya Duran. 2017. *Lexically Driven Errors: An Analysis of English Language Learners' Written Texts*. Dağyeli Verlag:Berlin.

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

Gaëtanelle Gilquin and Sylviane Granger. 2022. Using data-driven learning in language teaching. In *The Routledge Handbook of Corpus Linguistics*, pages 430–442. Routledge.

Lelia Glass. 2022. English verbs can omit their objects when they describe routines. *English Language & Linguistics*, 26(1):49–73.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Leslie MS Griffiths II. 2019. *COMPASS Placement Assessment and Student Attrition at a Community College*. Ph.D. thesis, Walden University.

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.

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

Muhammed Ali Khan. 2019. New Ways of Using Corpora for Teaching Vocabulary and Writing in the ESL Classroom. *ORTESOL Journal*, 36:17–24.

Young-Suk Grace Kim, Christopher Schatschneider, Jeanne Wanzek, Brandy Gatlin, and Stephanie Al Otaiba. 2017. Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and writing*, 30:1287–1310.

Jeffrey Klausman and Signee Lynch. 2022. From ACCUPLACER to Informed Self-Placement at Whatcom Community College: Equitable Placement as an Evolving Practice. *Writing placement in two-year colleges: The pursuit of equity in post-secondary education. The WAC Clearinghouse*, pages 59–83.

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. Detecting multiword expression type helps lexical complexity assessment. *arXiv preprint arXiv:2005.05692*.

Kristopher Kyle. 2019. Measuring lexical richness. In *The Routledge handbook of vocabulary studies*, pages 454–476. Routledge.

Éric Laporte. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer and Stella Markantonatou, editors, *Multiword expressions: In-sights from a multi-lingual perspective*, pages 143–186. Language Science Press, Berlin.

Ron Martinez. 2013. A Framework for the Inclusion of Multi-word expressions in ELT. *ELT journal*, 67(2):184–198.

Amy Mazzariello, Elizabeth Ganga, and Nikki Edgecombe. 2018. Developmental Education: An Introduction for Policymakers. *Education Commission of the States*.

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.

Daehyeon Nam and Kwanghyun Park. 2020. *I will write about*: Investigating multiword expressions in prospective students' argumentative writing. *Plos one*, 15(12):e0242843.

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.

Taha Omidian, Hesamoddin Shahriari, and Behzad Ghonsooly. 2017. Evaluating the Pedagogic Value of Multi-Word Expressions Based on EFL Teachers' and Advanced Learners' Value Judgments. *TESOL Journal*, 8(2):489–511.

Dolores Perin and Mark Lauterbach. 2018. Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, 28(1):56–78.

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.

Marco SG Senaldi, Debra A Titone, and Brendan T Johns. 2022. Determining the importance of frequency and contextual diversity in the lexical organization of multiword expressions. *Canadian*

Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 76(2):87–98.

Shelley Staples and Randi Reppen. 2016. Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, 32:17–35.

Haidee Thomson. 2017. Building speaking fluency with multiword expressions. *TESL Canada Journal*, 34(3):26–53.

A Phraseology Approach in Developmental Education Placement*

Miguel Da Corte[1][0000−0001−8782−8377] and Jorge Baptista[1,2][0000−0003−4603−4364]

[1] University of Algarve, Faculty of Human and Social Sciences
mlloveradacorte@gmail.com
[2] INESC-ID Lisboa, Human Language Technology Lab
jbaptis@ualg.pt

Abstract. This study focuses on an automatic classification task aiming at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. DevEd courses are designed to remediate and support students' communication skills in reading and writing before they can fully participate in college-level or college-bearing courses. This paper uses machine-learning methods to investigate the impact of considering multiword expressions (MWE) as entire tokens on the automatic classification task. Since many MWE are often non-compositional in meaning and constitute a large percentage of the textual units of many texts, they are likely to have a relevant role in the data representation of texts and, hence, improve subsequent classification task. Information is scarce regarding the tokenization of MWE and how this affects automatic placement. To this end, a random, balanced corpus of 186 sample texts (93 from each level) was used. Experiments compared the performance of a set of classifiers on the plain text corpus and on a version of the same corpus annotated for MWE. Results showed that using MWE as lexical features improved the classification accuracy by 8.1% above the baseline.

Keywords: Developmental education · machine-learning classification · multiword expressions · text mining.

1 Introduction and Objectives

The study of MWE continues to provide an opportunity for the enhancement of linguistic analysis. Several authors consider the lexical variety and repetition of MWE to be key in the process of language acquisition and mastery of language proficiency and fluency [9, 12, 15]. Understanding the lexical and syntactical patterns of community college students paves the way to understanding

* Research for this paper has been supported by University of Algarve, Language Sciences doctoral program, and by national funds through Fundação para a Ciência e a Tecnologia (Proj.Ref. UIDB/50021/2021).

Proceedings of the International Conference EUROPHRAS 2022, pages 79–86,
Malaga, Spain, 28-30 September, 2022.
https://doi.org/10.26615/978-954-452-080-9_011

issues related to language that impede effective communication both verbally and in writing. For community college students, the target population of this study, the use of these "prepackaged," "bundled" expressions facilitates communication and provides a sense of fluency, even if their writing skills require improvement to enter higher education. This is one of the main purposes of Development Education (DevEd) [4, 5, 16]. On the other hand, placement in DevEd courses often resorts to machine-learning based systems, such as ACCUPLACER[3], COMPASS[4], ACT[5], which use textual linguistic features to assist the placement strategies followed by higher education institutions. The COH-METRIX [10] [6] has been used to provide descriptive textual features that help capture Text Complexity and Readability, e.g., word count, paragraph length, word length; Text Easability; Referential Cohesion; Lexical Diversity; Connectives, Syntactic Complexity; Syntactic Pattern Density; Word Information; and Readability metrics (Flesch-Kincaid). These categories can be used as items for a linguistic and discourse representation to enrich corpora with linguistic data as evidenced in a previous study [1]. However, there is scarce information on how these systems deal with MWE (if at all), or at least what types of MWE are considered. The prevalent use of MWE in students' writing samples from entrance exams motivates the study of the impact of phraseology on machine-learning classification tasks pertaining to DevEd course placement, thus, succinctly defining the focus of this paper.

2 Related Work

Vocabulary in discourse and the connection of words within the larger discourse has been examined by [14], who particularly focused on how meaning is carried through *formulaic sequences*. Particular emphasis is placed on the formation of multiword units as a whole and common categories of these units such as compound words, phrasal verbs, fixed phrases, and lexical phrases are exemplified. The author asserts [14, p.101] that "language production stems from the precept that native speakers tend to use language that is formulaic in nature."

Understanding common categories of MWE requires a deep examination of their properties and the impact of these properties on NLP applications. According to [8], collocation and discontiguity phenomena are heavily exploited features, among the many properties MWE present, that have aided parsing tasks and, thus, enhanced syntactic analysis [3]. Studying the impact of discriminated tokens versus *single syntactic constituents*, as coined by [8], on the structural and functional aspects of developing writers' skills can aid in a more accurate placement of students in DevEd [8].

[3] `https://www.accuplacer.org/` (Last access: July 12, 2022; all URL in this paper were check on this date.)
[4] `https://www.compassprep.com/practice-tests/`
[5] `https://www.act.org`
[6] `http://141.225.61.35/CohMetrix2017/`

The incidence of multiword expressions was studied by [11] in 746 argumentative essays with 121,638 tokens from Korean-university students with the goal of exposing reocurrent structural sentence patterns and their frequency. These patterns were compared to those exhibited by American-university students in two large corpora (LOCNESS; MICUSP), where the incidence of phrased-based expressions exhibited by L2 compared to the incidence of noun phrases by L1 emerged as a theme. Additionally, L2 student showed a wider variety of MWE in their writing, suggesting that the use of these lexically-bound expressions facilitates communication among developing writers by adding a level of functionality to the writing process. Even though this study focused on L1 and L2 students, it supports the investigation of "MWE across proficiency levels or novice and professional academic writing [...] in an academic context." [11, p.11]

The ongoing variability of MWE, particularly verbal ones, and the challenges it poses for automated machine learning identification tasks are recognized by [13]. The authors focused on strategies to improve the identification of verbal MWE (VMWE) and used the PARSEME[7] corpora as a starting point. They completed three tasks comprised of a training and development phase, a prediction phase, and an evaluation phase all aided by a simple [candidate VMWE potential] extraction of filtering (*Seen*2020) techniques for precision [13]. Promising results were evidenced in boosting the identification of global MWE by using a combination of morphosyntactic filters. Suggestions for further work emphasize the representation of "VMWE as multisets of (lemma-POS) pairs rather than lemma and POS multisets separately." [13, p.3342]

A framework to further investigate the lexical and syntactical features of MWE is provided by [8, 11, 13, 14], among others. However, it is unclear whether automatic placement classification systems, i.e., ACCUPLACER, COMPASS, and ACT, take MWE into consideration or at least what types of MWE are considered. To the best of our knowledge, these systems provide a holistic score of writing samples based on the purpose and focus of the essay; organization and structure; development and support of ideas; sentence variety and style; accurate use of Standard Written English; and how one communicates and connects ideas while addressing a topic [2]. However, [16] emphasizes the need to improve college readiness by more accurately assessing students' performance in writing tasks. We aim at detecting MWE and MWE types to aide in the assessment of lexical features towards language proficiency.

3 Methods

To assess the impact of MWE in our classification task, a small corpus of 186 essays was collected from community college students, in the State of Oklahoma, U.S.A., placed in DevEd after completing their writing placement exam during years 2020-2021. The sample texts, consisting of a short multiparagraph essay of 300-600 words, prompted by a short excerpt with guiding questions, were

[7] `https://typo.uni-konstanz.de/parseme/`

retrieved from the ACCUPLACER platform of the higher education institution where this study took place. The corpus was then balanced for placement level (93 essays from each level: Level 1 and Level 2), and for sample size (keeping only up to 100 tokens per text). All the essays addressed the same topics.

Two experiments were devised, where the Classification Accuracy (CA) performance of a set of classifiers of the raw corpus and on a version of the same corpus annotated for all types of MWE was compared. We first devised a classification criteria for identifying MWE. Then, we independently annotated the corpus and compared the annotations in order to validate the selected MWE. The ORANGE[8] data-mining toolkit [6] was chosen for analysis and modeling since it provides, in a practical way, several NLP tools (mainly for preprocessing and data representation) within the same machine-learning (ML) toolkit and a large set of ML algorithms and data visualization tools.

3.1 Corpus Annotation and Classification Guidelines

In order for the Orange data-mining toolkit to tokenize a MWE as a single token, its elements had to be joined in the input text (an underscore was used to this end). This task was done manually. One of the reasons for this has to do with the large number of typos and spelling errors found in the essays, which we could only ignore with a manual inspection of the text. For the manual identification of MWE, the set of linguistic criteria adopted derived from the literature [3]. The MWE categories of [7] and the caveats by [8] were particularly insightful. Based on these criteria, the annotation of MWE in the corpus included multiple categories, based on traditional classification guidelines and lists available in dictionaries, e.g., **compound nouns**: *golden rule*, *refresher course*; **compound adjectives**: (stay) *tall and straight*, (be) *on* [one's] *feet*; *strong minded*; **compound conjunctions and prepositions**: *as long as*, *in spite of*; **compound adverbs**: *by the way*, *back in the old day*; **verbal idioms**: *see the bigger picture*, *knock at your door*; *weigh the pros and cons*; **phrasal verbs**: *fit in*, *mess up*; **support verb constructions**: *have no clue*, *make mistakes*; among others. In the case of **nouns**, these also included named entities, both person names (*Colin Powell*), locations (*Guatemala City*), and organizations (*United Nations*).

For the annotation proper, the MWE were joined following these guidelines: Sequential strings of words were just joined by an underscore (*phone_call*). Productive inserts, as in phrasal verbs, were marked by a double underscore and the inserted word permuted (*bring you down* → *bring__down you*). Long-distance elements of the same MWE were joined in the first element, and the slot left void marked by a hashtag '#': The more *success <sic> you are in your job* the more *you will make* → *The_more_the_more success you are in your job # you will make*. Prepositions introducing complements of predicative elements (adjectives, nouns, and verbs, including phrasal verbs) are not taken into account, e.g. *to get_out of something* where *of something* is just a complement of the phrasal verb *get_out* (joined). At the end, 1,260 MWE were annotated.

[8] `https://orangedatamining.com/`

3.2 Data Processing and Modeling

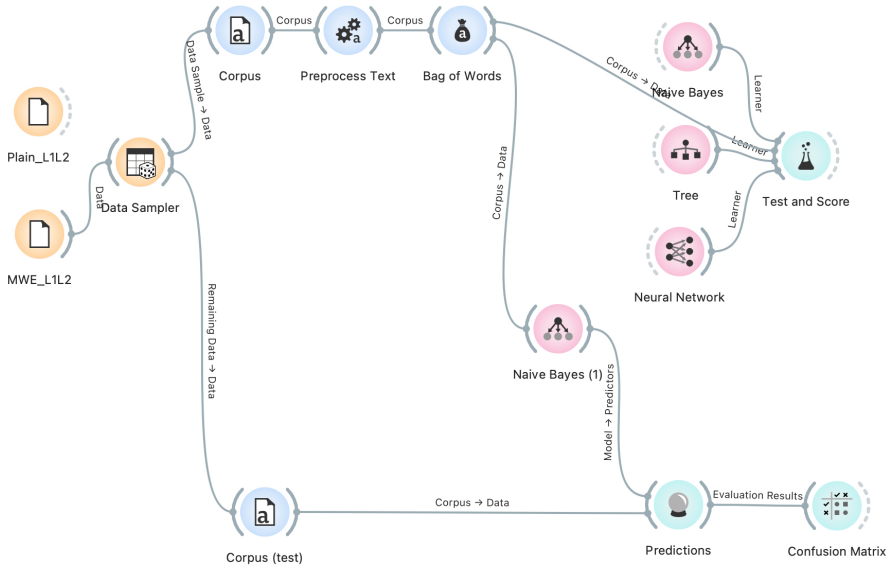


Fig. 1. Orange Text Mining Environment

The workflow adopted for this study is shown in Fig. 1. The DATASAMPLER widget was used to partition the data for a 3-fold cross-validation, leaving 2/3 for training and 1/3 for testing purposes. In the PREPROCESSING stage, basic tokenization options were selected, capturing both words and punctuation as tokens. Experiments were then carried out introducing different PART-OF-SPEECH (PoS) taggers. The data representation chosen was BAG-OF-WORDS (using the count of term frequency). For the training step, and to assess the models, the TEST&SCORE widget was used. The models used were Naive Bayes, Decision Tree (forward pruning) and Neural Network (multi-layer perceptron with back propagation). The data representation and the learning models were chosen because they had already proven to perform well with this type of data in previous experiments. For the PREDICTIONS (testing) step, the best performing model in each configuration setting was assessed.

3.3 Experiments

Table 1 shows the experimental settings and the breakdown of the results. In these experiments, we compared the non-annotated corpus (PlainText) with the corpus where MWE had been joined (Text-MWE). Since the corpus is small, the training partition corresponds to 2/3 of the corpus and the remaining is

Table 1. Experimental settings and results.

Experiment	Description	NB	NN	TR	best model	
Baseline	PlainText;Tok:W&P. noPoS.BoW	**0.685**	0.629	0.669	NB	0.742
Experiment 01	PlainText; idem+PoS:TB-ME	0.669	**0.718**	0.669	NN	0.677
Experiment 02	PlainText; idem+PoS:AP	0.613	0.661	**0.685**	TR	0.726
Experiment 03	Text-MWE;Tok:W&P. noPoS.BoW	0.707	**0.740**	0.553	NN	**0.823**
Experiment 04	idem+PoS:TB-ME	0.573	0.661	**0.667**	TR	0.742
Experiment 05	idem+PoS:AP	0.540	0.637	**0.653**	TR	0.758

left for testing. The data representation mode was Bag-of-Words with a simple frequency count. Results are provided using the Classification Accuracy metrics (CA). The names of learning models selected for the study are shortened to NB (Naive Bayes), Neural Networks (NN) and Tree (TR). The values reported for each individual model concern the Test&Scoring (training) step. In the rightmost columns, the CA values in the testing step are reported to the best performing model in the previous training step. In the description of the experiments, the two PoS-taggers have been shortened: Treebank - Maximum Entropy (TB-ME) and Average Perceptron (AP). Otherwise, 'noPoS' indicates no PoS-tagging was applied.

We first defined a baseline, using a basic configuration for the preprocessing step. The Word&Punctuation tokenization option was chosen, because it keeps punctuation as is, which might be a factor of placement classification. This is due to the fact that the use of punctuation is one of the writing skills not yet entirely mastered by the students and which is a topic that is addressed by the DevEd courses. No difference was found when the models were trained with different combinations of other tokenization configurations, namely just using white-space or just the words. No PoS-tagger was used in this baseline. For Experiments 1-2, the same baseline configuration was used but now adding each PoS-tagger in turn. For Experiment 3, the initial baseline configuration was enhanced by using the same corpus but with its MWE joined as single tokens. No PoS-tagger was used here. In Experiments 4-5, the same enhanced configuration was used, with joined MWE, but now adding each PoS-tagger.

4 Results and Discussion

Overall, the results from the experiments showed that adding MWE information to single-word tokenized texts improves the classification up to a clear 8.1% above the baseline (Experiment 03). The preprocessing tokenization options available with Orange did not seem to affect results too much. On the other hand, PoS-tagging underperforms the baseline, especially the AP (Experiment 02). With the MWE annotated text, however, AP performs better (Experiment 05) than TB-ME (Experiment 04). When one compares the results from the best performing model in the training step with its results in the testing step, the testing step results usually outperform the training, except in Experiment 01, where testing

results were 4.1% less than training. Otherwise, differences vary from 4.1% (Experiment 02) to 10.5% (Experiment 05). While using the plain text corpus in the training step, and though results are not very different from the baseline, no model proved to outperform the other two, as each in turn occurs as the best model. Using the text with MWE, however, showed that Neural Networks was the best model (Experiment 03), but that Tree outperformed the other two models when PoS-tagging was added to the configuration.

Although the corpus is small, the size of MWE found in it corresponds to 6.75% of the words in the corpus, a non-neglegible percentage. Still, results seem to signal the importance of using information on MWE in this task. This is in line with the literature findings on similar scenarios [7]. A proper PoS-tagging of MWE, combined with an appropriate classification of MWE types [9] may contribute to improving results and providing more robust insights on DevEd students' writing patterns.

5 Conclusion and Next Steps

This study focused on an automatic classification task aimed at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. We investigated and presented experimental results on the impact that multiword expressions (MWE), considered as entire tokens, have on an automatic classification task using machine-learning methods. The classification was based on linguistic features derived from small corpus of texts written by native English speakers at the onset of their higher education journey. Issues of orthography, punctuation, lexical usage such as slang, grammatical issues such as agreement, semantic issues, and discursive aspects of the text have been previously explored. Now, the inclusion of MWE provided promising results as they constitute a large percentage of the textual units of many texts and are likely to have a relevant role in the representation of the information in texts. An extended corpus will be required to confirm the hypothesis presented in this preliminary study.

A quantitative evaluation to verify the usefulness of including MWE information was presented and discussed, including the benefits of using different PoS-taggers and the degree to which these tools affect the classification accuracy. Although, at this point in the research study, we are not concerned with the automatic identification of MWE in DevEd students' writing, we are committed to developing, first, a more precise MWE scheme for annotating writing samples to further gain insights into linguistic issues that prevent effective communication for this student population.

[9] The list of MWE is available at: `https://www.researchgate.net/project/Linguistic-Aspects-of-Developmental-Education`

References

1. Abba, K.A.: Community college students' writing: Lexical, syntactic, and cohesion differences in L1, L2, and Generation 1.5 students and examining knowledge of the writing process. Ph.D. thesis, Texas A&M University, Graduate and Professional Studies (2015)
2. Board, C.: ACCUPLACER program manual. The College Board New York (2018)
3. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword Expression Processing: A Survey. Computational Linguistics **43**(4), 837–892 (12 2017). https://doi.org/10.1162/COLI_a_00302, `https://doi.org/10.1162/COLI\_a\_00302`
4. Cormier, M., Bickerstaff, S.: Research on developmental education instruction for adult literacy learners. The Wiley Handbook of Adult Literacy pp. 541–561 (2019)
5. Darkenwald-DeCola, J.A.: 'In College, I'm the One People Go To': Lessons from Successful Developmental Literacy Students About the Transition to College-Level Courses Across Disciplines. Ph.D. thesis, Rutgers The State University of New Jersey, School of Graduate Studies (2021)
6. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research **14**, 2349–2353 (2013), `http://jmlr.org/papers/v14/demsar13a.html`
7. Kochmar, E., Gooding, S., Shardlow, M.: Detecting multiword expression type helps lexical complexity assessment. arXiv preprint arXiv:2005.05692 (2020)
8. Laporte, E.: Choosing features for classifying multiword expressions. In: Sailer, M., Markantonatou, S. (eds.) Multiword expressions: In-sights from a multilingual perspective, pp. 143–186. Language Science Press, Berlin (2018). https://doi.org/10.5281/zenodo.1182597
9. Martinez, R.: A framework for the inclusion of multi-word expressions in elt. ELT journal **67**(2), 184–198 (2013)
10. McNamara, D.S., Ozuru, Y., Graesser, A.C., Louwerse, M.: Validating CoH-Metrix. In: Proceedings of the 28th annual Conference of the Cognitive Science Society. pp. 573–578 (2006)
11. Nam, D., Park, K.: *I will write about*: Investigating multiword expressions in prospective students' argumentative writing. Plos one **15**(12), e0242843 (2020)
12. Omidian, T., Shahriari, H., Ghonsooly, B.: Evaluating the Pedagogic Value of Multi-word Expressions based on EFL Teachers' and Advanced Learners' Value Judgments. TESOL Journal **8**(2), 489–511 (2017)
13. Pasquer, C., Savary, A., Ramisch, C., Antoine, J.Y.: Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3333–3345 (2020)
14. Schmitt, N., Schmitt, D.: Vocabulary in language teaching. Cambridge University Press (2020)
15. Thomson, H.: Building speaking fluency with multiword expressions. TESL Canada Journal **34**(3), 26–53 (2017)
16. Zachry Rutschow, E., Edgecombe, N., Bickerstaff, S.: A brief history of developmental education reform (Oct 2021), `https://postsecondaryreadiness.org/research/history-developmental-education-reform/`

ETIQUETAJE DE EXPRESIONES MULTIPALABRA EN ENSAYOS ESCRITOS POR NATIVOS Y NO NATIVOS DE ESPAÑOL EN UN CURSO DE DESARROLLO DE GRAMÁTICA Y COMPOSICIÓN

Miguel Da Corte

Universidade do Algarve, Faculdade de Ciências Humanas e Sociais,
Campus de Gambelas, P-8005-139 Faro, Portugal
A74072@ualg.pt

Jorge Baptista

INESC-ID Lisboa, HLT Lab, Rua Alves Redol, 9, P-1000-029 Lisboa, Portugal
jbaptis@ualg.pt

Multiword expression tagging of Spanish native and non-native speakers' written essays in a grammar and composition developmental course

Abstract: The literature on second language learning posits that there are significant differences between the use of multiword expressions (MWE) by native speakers (NS) and non-native speakers (NNS). Furthermore, it considers that levels of language proficiency can be estimated on the basis of the use of these expressions. This paper analyses the written production from a corpus of essays written by native (16 essays, 5839 words) and non-native Spanish speakers (25 essays, 7767 words) enrolled in a course focused on the development of orthographic, grammatical, lexical, semantic, and discursive skills in Spanish. This is a required course for students pursuing a certification in Translating or Interpreting (Spanish/English) in the educational setting where the study took place. The corpus was manually tagged by two linguists. The classification scheme used was inspired by other schemes found in the literature and built for similar purposes. The results show that, in general, the distribution of MWE types found in the NS and NNS partition of the corpus was not very different (Pearson correlation: 0.894). However, interesting differences were found between the categories of verbal idioms and noun constructions. Though the corpus is too small for more significant conclusions to be drawn, it is possible

Romanica Olomucensia 35/1 (2023): 23-40, doi: 10.5507/ro.2023.003 (CC BY-SA 4.0)

to point out that different types of MWE are unevenly distributed among the native speakers' and non-native learners' written production material, and some categories may be a clearer indicator of near-native-speaker proficiency.

Keywords: multiword expressions; language proficiency; classification level; machine-learning models; developmental education courses (in Spanish)

Resumen: La literatura sobre el aprendizaje de una segunda lengua postula que existen diferencias significativas entre el uso de expresiones multipalabra (EMP) por parte de hablantes nativos (HN) y no nativos (HNN). Además, considera que los niveles de competencia lingüística pueden estimarse a partir del uso de dichas expresiones. En este trabajo se analiza la producción escrita de un corpus de ensayos escritos por hablantes nativos (16 ensayos, 5.839 palabras) y no nativos de español (25 ensayos, 7.767 palabras) matriculados en un curso centrado en el desarrollo de las habilidades ortográficas, gramaticales, léxicas, semánticas y discursivas en español. Este es un curso de matriculación obligatoria para los estudiantes que aspiran a obtener el título de traductor o intérprete (español/inglés) en el centro educativo donde se realizó el estudio. Dos expertos lingüistas etiquetaron de forma manual el corpus del estudio. El esquema de clasificación utilizado se inspiró en otros esquemas encontrados en la literatura y construido con fines similares. Los resultados mostraron que no se encontraron mayores diferencias (correlación de Pearson: 0,894) en la distribución de los tipos de EMP en el análisis del corpus de HN y HNN. Sin embargo, se observaron diferencias interesantes en algunas categorías, concretamente entre las expresiones fraseológicas verbales y los sustantivos comunes compuestos. Aunque el corpus es pequeño para llegar a conclusiones más relevantes, cabe destacar que los diferentes tipos de EMP se distribuyen de forma desigual en los textos escritos por los hablantes nativos y no nativos y que algunas categorías son un indicador más claro de un nivel de competencia más cercano al nivel de dominio del idioma materno.

Palabras clave: expresiones multipalabra; competencia lingüística; nivel de clasificación; modelos de aprendizaje automático; cursos para el desarrollo de habilidades (en español)

1. Introducción

El uso de expresiones multipalabra (EMP) ha sido motivo de continuas investigaciones debido a su idiosincrasia, complejidad y al enfoque usado para su estudio (García-Page 2008; Corpas Pastor 2017; Pasquer *et al*. 2020). A pesar de que existen diferentes criterios para la evaluación, señalización y estandarización de dichas expresiones (Farahmand *et al*. 2015), los rasgos de coocurrencia e incidencia se presentan como criterios básicos para el esclarecimiento de la función y el valor que las EMP añaden a la producción de textos escritos (Savary *et al*. 2019).

En un análisis para predecir el nivel de competencia lingüística[1] en inglés de estudiantes franceses, y tomando como referencia el Marco Común Europeo de Referencia para las Lenguas[2] (MCER), Arnold *et al*. (2018) examinaron un modelo predictivo en el que el nivel de competencia lingüística se establecía como una función dependiente del grado de complejidad de diferentes estructuras lingüísticas. Todos los datos pertenecientes a los textos no procesados fueron extraídos de la base de datos

[1] Se adopta el término *competencia lingüística* como traducción al español del concepto *proficiency*, en inglés.

[2] <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/>. Todas las páginas web en este documento se consultaron por última vez el 7 de marzo de 2023.

EFCAMDAT[3] (*EF Cambridge Open Language Database*). Se utilizó un total de 41.626 textos, correspondientes a 128 temas de redacción, de una población de 7.695 estudiantes franceses. Basándose en nueve 9 indicadores de complejidad lingüística (por ejemplo, número de palabras, número de frases, frases verbales, frases coordinadas y frases nominales complejas) y utilizando la herramienta *LE Syntactic Complexity Analyzer*[4] (L2SCA), los resultados revelaron que el número de *tokens* y el tipo de palabra desempeñan un papel fundamental en el cálculo del nivel de competencia lingüística. Concretamente, «la habilidad de generar expresiones y frases complejas y usar correctamente estructuras avanzadas, tales como oraciones relativas sin pronombre relativo, son indicadores de estudiantes con un nivel avanzado en la lengua meta» (Arnold *et al*. 2018: 4; traducción propia).

La producción de estas expresiones y frases complejas ha sido objeto de interés en el campo de Procesamiento del Lenguaje Natural (PLN) (en inglés, *Natural Language Processing*). Parte del interés reside en: (i) la relevancia de estas expresiones en la clasificación (y calificación) automática de ensayos escritos «con otras características clásicas como longitud de las EMP, recurso de clasificación (*graded resource*), vecinos ortográficos (*orthographical neighbours*), partes de la estructura gramatical de una oración, morfología, relaciones de dependencia, tiempo verbal, desarrollo del lenguaje y coherencia» (Wilkens *et al*. 2022: 62; traducción propia); (ii) el efecto que ciertas tareas o trabajos escritos tienen en la precisión de las estructuras lingüísticas usadas por aprendices de una segunda lengua y para lo cual se necesitan herramientas PLN para asegurar la exactitud de los resultados (Alexopoulou *et al*. 2017); (iii) la pedagogía que se utiliza en la enseñanza de la escritura de una segunda lengua y cómo esta influye en los índices de desarrollo de competencia lingüística y dependencia en EMP (El-Dakhs *et al*. 2022; Esfandiari & Mohammad 2022).

El uso de expresiones y frases complejas, en su mayoría, se ve condicionado por el tema de redacción o las instrucciones proveídas, lo que la literatura advierte debe monitorearse para así evitar una distribución sesgada de dichas expresiones. Basándose en el número promedio de palabras por frase, la longitud promedio de estas y la relación de frases subordinadas con respecto al total de frases utilizadas, Alexopoulou *et al*. (2017) sostienen que la longitud de las frases aumenta con el nivel de competencia y que la complejidad lingüística se ve afectada por la complejidad de la tarea a desempeñar.

A partir de estos rasgos lingüísticos, el objetivo que nos proponemos consiste en experimentar con técnicas de PLN para realizar un análisis comparativo de textos escritos por hablantes nativos y no nativos de español. Este análisis permitirá estimar y confirmar el nivel de competencia lingüística de los hablantes no nativos, teniendo en cuenta la diversidad de EMP usadas en sus producciones escritas. Según nuestros conocimientos, ningún estudio sistemático sobre el uso de la EMP por parte de hablantes nativos se ha correlacionado con una evaluación sobre el nivel de competencia lingüística de hablantes no nativos a nivel académico. Esta investigación

[3] <https://philarion.mml.cam.ac.uk/>.
[4] <https://sites.psu.edu/xxl13/l2sca/>.

representa un primer paso hacia ese objetivo. Al mismo tiempo, procuramos identificar los rasgos lingüísticos que son más predictivos del nivel de competencia escrita de los hablantes no nativos versus los hablantes nativos y si ciertas EMP –tipos y *tokens* con combinaciones de dos o más palabras– son usadas por uno de los grupos con mayor frecuencia.

La necesidad concreta de contrastar y analizar el uso de EMP entre hablantes nativos y no nativos surge del hecho de que el dominio de estas expresiones, ya catalogado como pertinente para los niveles superiores de competencia según el MCER (Consejo de Europa 2002), se asocia, según estudios previos (Dahunsi & Ewata 2022; Hinkle 2023), con altos niveles de competencia lingüística. Apoyando esta afirmación, en un análisis sobre el uso de EMP (Da Corte & Baptista 2022a), realizado en el mismo contexto que aquí se presenta, se analizó la producción escrita de hablantes nativos de inglés (L1) que participaron en un curso de dos niveles (nivel 1 y 2) de desarrollo de gramática y composición como requisito indispensable para ser admitidos en un programa académico. Los experimentos realizados compararon el efecto sobre un conjunto de clasificadores de un corpus sin anotaciones y en una versión del mismo corpus anotada con EMP. Los resultados mostraron que la identificación de las EMP, como características léxicas, mejoró la precisión de la clasificación de textos en un 8,1 %, en cuanto a los niveles (nivel 1 y 2) del curso de gramática y composición, por encima del experimento base (sin anotación) y consecuentemente la ubicación en el nivel correspondiente en dicho curso. Ya que el contexto es el mismo y la expectativa general se ajusta al marco teórico de este estudio, se seleccionó como indicador principal de este estudio las EMP.

En cuanto a la organización de este documento, primero, se muestra, en la sección de métodos, una descripción detallada del contexto institucional y los participantes de este estudio (Sección 2.1; 2.2). Se incluye también información sobre la constitución del corpus (2.3), así como los tipos de EMP seleccionadas para el etiquetaje del corpus y el esquema de anotación (2.4). Los experimentos con herramientas de aprendizaje automático se presentan de manera resumida (2.5) con los resultados y algunas reflexiones (3). Las conclusiones y trabajo futuro se presentan en la última sección (4).

2. Métodos

En esta sección se incluye: el contexto institucional, los participantes con el nivel de competencia lingüística correspondiente, el corpus utilizado, los tipos de EMP y esquema de anotación empleados y los respectivos experimentos con herramientas de aprendizaje automático.

2.1. Contexto institucional y participantes

Un total de veintidós estudiantes de un instituto de educación superior (Oklahoma, Estados Unidos) participaron en este estudio. Estos estudiantes estaban matriculados en un programa (bilingüe) de entrenamiento intensivo de traducción e interpretación (inglés-español; español-inglés). De estos veintidós estudiantes, quince manifestaron

tener el inglés como lengua materna y siete el español, siendo esta la lengua que usan con más frecuencia.

Ambos grupos participaron en un curso de dieciséis semanas de español como segunda lengua que tenía como objetivo principal el desarrollo de aspectos ortográficos, gramaticales, semánticos y discursivos. Este curso es denominado Estudio Intermedio de Gramática y Composición en Español y se imparte completamente en español, la lengua meta. Para los hablantes nativos de español, la participación en este curso les permite afianzar el conocimiento en cuanto a su lengua materna o simplemente confirmar sus niveles de competencia. Este curso es uno de los requisitos en el plan de estudios de Traducción e Interpretación, además de ser un elemento indispensable para ser técnico superior (primer paso en la obtención de una licenciatura). Para los hablantes no nativos de español, este curso permite corregir y desarrollar algunos patrones léxico-sintácticos y discursivos que llevan a un aumento del nivel de competencia lingüística en una segunda lengua. Cabe destacar que la participación en este curso, para los hablantes no nativos, es posible después de completar satisfactoriamente tres semestres de español – Niveles 1 al 3.

2.2. Nivel de competencia de los participantes

Según la tabla de convalidaciones del MCER (Consejo de Europa 2002) y el Consejo Americano para la Enseñanza de Idiomas Extranjeros[5] (*ACTFL*, por sus siglas en inglés), en relación con la escala de competencia lingüística, los estudiantes no nativos muestran un nivel de competencia equivalente al B1/B2. Este nivel de competencia se estima empleando dos sistemas de evaluación: uno formativo y otro sumativo. La evaluación formativa incluye la redacción de un párrafo sobre un tema en particular y un ejercicio de comprensión oral, con un fragmento de texto pregrabado, en la que los participantes responden a 5 preguntas de selección múltiple o verdadero/falso. La evaluación formativa se aplica al final de cada lección (6 en total), en cada nivel. La evaluación sumativa consiste en una prueba oral, con preguntas y respuestas, usando ciertos temas de discusión que únicamente son compartidos en el momento de la prueba (que se aplica solo al final del semestre). La escala de evaluación oral se enfoca en cuatro aspectos principales: número de funciones lingüísticas, contexto y contenido, precisión y tipo de discurso (ACTFL 2016); (Alpine Testing Solutions 2020).

Como parte del programa de estudio, los hablantes nativos y los no nativos demuestran y desarrollan sus habilidades de producción escrita según cuatro enfoques discursivos: descripción, narración, información y opinión. Cada ensayo tiene un tema abstracto en particular, por ejemplo: (i) relaciones personales en la sociedad global de hoy; (ii) experiencias personales en ciudades grandes, cosmopolitas en comparación con ciudades más pequeñas y poco desarrolladas; (iii) impacto de la tecnología en la vida cotidiana; y (iv) la convivencia en familia. Estos temas, indicados aquí a modo de ejemplo, fueron propuestos a los participantes y sobre ellos se realizaron los ensayos que constituyen el corpus objeto de este estudio. Las producciones escritas de ambos grupos se caracterizan por el uso y dominio de diferentes tiempos verbales, así como por otros rasgos lingüísticos de interés, particularmente el uso de EMP.

2.3. Constitución del corpus

El corpus utilizado para este análisis tiene un total de 81.466 caracteres que provienen de la colección de textos escritos producidos por los hablantes nativos y no nativos de español en el curso de gramática y composición. Se obtuvo una muestra de dos ensayos por estudiante durante dos períodos claves en el semestre en el que se desarrolló el curso: uno al principio del semestre (antes de recibir formación sobre contenidos gramaticales y ortográficos) y el segunda después de cinco semanas. En cuanto al contenido de las muestras, los ensayos de los hablantes nativos contienen aproximadamente 2.140 caracteres, con una desviación media de 1.203; en comparación, los ensayos de los no nativos contienen un 9 % menos en el promedio de caracteres usados (1.889) y una desviación media de 964. El tamaño de los ensayos escritos por parte de los hablantes nativos oscila entre los 383 y 4.711 caracteres, mientras que el número de caracteres en los ensayos de los hablantes no nativos oscila entre 691 y 4.647.

Estas cifras, que se presentan de manera resumida en la Tabla 1, demuestran que el tamaño de los ensayos de los hablantes nativos y no nativos es comparable, pero confirman la necesidad de tener un corpus equilibrado en cuanto al número de caracteres por unidad de muestreo, por nivel y temas presentados en los ensayos.

Indicador	HN	HNN	Corpus
Número de caracteres (mínimo)	383	691	383
Número de caracteres (máximo)	4.711	4.647	4.711
Número total de caracteres	**34.247**	**47.219**	**81.466**
Media	2.140	1.889	1.987
Desviación media	1.203	964	1.056

Tabla 1. Número de caracteres, media y desviación media de los ensayos escritos por los hablantes nativos (HN) y hablantes no nativos (HNN)

Dos lingüistas, con experiencia en tareas de anotación de corpus y dominio en los conceptos de EMP utilizados en este estudio, etiquetaron manualmente el corpus. Dado el pequeño tamaño del corpus, la anotación se llevó a cabo de forma independiente y las discrepancias existentes se discutieron para llegar a una delimitación y clasificación consensuadas.[5] Los elementos de cada una de las EMP se enlazaron con un carácter de subrayado o barra baja ('_'), para así ser considerados y tratados como una sola palabra gráfica (*token*, en inglés) en la herramienta de aprendizaje automático, Orange[6]. Esta herramienta se utiliza en experimentos de clasificación automática y en procesos de minería de datos. Las etiquetas siguen los lineamientos diseñados y aplicados con anterioridad a un corpus en inglés de estudiantes anglohablantes que desarrollan sus habilidades en su lengua materna (Da Corte & Baptista 2022a).

[5] Dadas las condiciones descritas, el cálculo del acuerdo entre anotadores no se llevó a cabo.
[6] <https://orangedatamining.com/>.

Las categorías de EMP (con sus respectivos códigos) delimitadas y empleadas en el etiquetaje manual del corpus fueron las siguientes: adverbios (ADV); locuciones conjuntivas (CONJ); sustantivos comunes compuestos (N); entidades nombradas (EN); preposiciones (PREP); pronombres (PRON); construcciones con verbo soporte (SVC); y, por último, expresiones fraseológicas verbales (VIDIOM).

2.4. Tipos de EMP y esquema de anotación

Las etiquetas aplicadas al corpus de los hablantes nativos y no nativos, en preparación para el análisis sobre la incidencia de EMP, se agrupan bajo la categoría de patrones lexicales y sintácticos. Estos patrones contribuyen con la construcción de frases y secuencias de palabras que, al combinarse con otras frases o palabras, definen o moldean el nivel discursivo del escritor y, por consiguiente, su nivel de competencia. El uso e incidencia de las EMP contribuyen a un análisis más holístico de ensayos escritos (Siyanova-Chanturia & Spina 2020), así como también resaltan, con mayor precisión, el desarrollo de las habilidades escritas en cuanto al nivel de competencia.

Debido a la variedad de las EMP, se utilizó como marco referencial para la clasificación de dichas expresiones una combinación de criterios formales, sintácticos y semánticos ya establecidos por otros autores como: Laporte (2018), quien presenta sugerencias en cuanto a la clasificación de construcciones con verbo soporte; Pasquer *et al*. (2020), quienes mencionan las expresiones fraseológicas verbales y describen su variabilidad; Nam & Park (2020) y Kochmar *et al*. (2020), quienes: (i) incluyen preposiciones, entidades nombradas, locuciones conjuntivas y sustantivos comunes compuestos en su lista de tipos de EMP; (ii) presentan sugerencias en cuanto a la construcción de modelos de aprendizaje automático para la clasificación de textos (por nivel), teniendo en cuenta la proporción e incidencia de las EMP que estos modelos reconocen en las fases de entrenamiento y prueba.

Los tipos de EMP, con sus respectivas definiciones, etiquetas, enlaces ('_') y algunos ejemplos[7] (incluyendo EMP que son parcialmente fijas), se presentan a continuación:

- Adverbio (ADV): EMP que funciona como adverbio simple. Ejemplos de adverbios empleados por los hablantes nativos incluyen: *casi_siempre*; *con_más_gusto_que_nunca*; *con_el_pasar_del_tiempo*; *al_contrario*; *en_solo_unas_palabras*; *en_la_actualidad*; *desde_mi_perspectiva*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_el_pasado*; *por_primera_vez*; *en_cualquier_lugar*; *por_supuesto*; *en_el_mundo_de_hoy*; *cara_a_cara*; *a_menudo*; *con_mucho_gusto*.
- Locución conjuntiva (CONJ): EMP usada como conjunción enlazada con un verbo. Ejemplos de conjunciones empleadas por los hablantes nativos incluyen: *a_medida_que*; *si_no*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_vez_de*; *antes_de_que*; *debido_a*; *así_que; como_si*.
- Sustantivo común compuesto (N): EMP usada como sustantivo. Ejemplos de sustantivos comunes empleados por los hablantes nativos incluyen: *roles_de_género*;

[7] La lista completa de EMP sin registros duplicados puede consultarse en Da Corte & Baptista (2022b).

nuevas_generaciones; *redes_sociales*; *señal_de_afecto*; *salud_mental*; *pista_de_hielo*; *pensamiento_crítico*; *lenguas_nativas*; *programas_de_noticias*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *cultura_de_la_soltería*; *relaciones_personales*; *habilidades_sociales*; *tiempo_libre*; *riesgo_financiero*; *palabras_sabias*; *almas_gemelas*; *casas_embrujadas*; *medios_de_interacción_social*.

- Entidad nombrada (EN): EMP que designa una entidad extralingüística y hace referencia a personas, organizaciones, lugares y eventos (Erdmann *et al.* 2019). Ejemplos de entidades nombradas empleadas por los hablantes nativos incluyen: *Guglielmo_Marconi* (persona); *Cristo_Rey* (lugar); *Pew_Research_Center* (organización); *Tulsa_Race_Massacre* (evento). En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *Tacos_Don_Francisco* (organización); *Acción_de_Gracias* (evento); *Nueva_York* (lugar).
- Preposición (PREP): EMP que funciona como una preposición enlazada con un sustantivo. Ejemplos de preposiciones empleadas por los hablantes nativos incluyen: *a_diferencia_de*; *por_culpa_de*: *en_frente_de*; *a_través_de*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_lo_que_respecta_a*; *en_torno_a*; *al_otro_lado_de*; *en_lugar_de*; *después_de*.
- Pronombre (PRON): EMP que funciona como pronombre. Entre los ejemplos de los pronombres empleados por los hablantes nativos y los no nativos, destacan los siguientes: *todo_el_mundo* (indefinido); *el_uno_al_otro* (complemento de eco); *sí_mismas*; *entre_sí* (recíprocos).
- Construcción con verbo soporte (SVC) o verbos leves (en inglés, *light verb constructions*): EMP que resulta de la combinación de un verbo con un sustantivo actuando como un predicado semántico, de la misma manera que un verbo simple o adjetivo, independientemente de la existencia de una nominalización (sustantivo-adjetivo; sustantivo-verbo). Este sustantivo predicativo selecciona argumentos y determina las propiedades sintácticas (estructurales) y semánticas distribucionales de sus oraciones base, así como las transformaciones que permite. En este trabajo, se siguió la perspectiva del Léxico-Gramática (Gross 1996) presentada recientemente por Fotopoulou *et al*. (2021) y Baptista *et al*. (2022) para la definición de las construcciones con verbo soporte. Ejemplos de construcciones con verbo soporte empleadas por los hablantes nativos incluyen: *juega_un_papel*; *tener_cuidado*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *asumir_el_compromiso*; *tiene_celos*; *dar_apoyo*; *hacer_diligencias*; *ofrecer_paz*; *guardo_gratos_recuerdos*; *pasar_#_tiempo*; *harás_conexiones*.[8]
- Expresión fraseológica verbal, frecuentemente idiomática (VIDIOM): EMP que resulta de la combinación de un verbo con por lo menos un argumento (sujeto; objeto) y que juntos tienen fuertes restricciones combinatorias. Estas expresiones suelen ser idiomáticas, es decir, su significado global no es composicional. Ejemplos de expresiones fraseológicas verbales empleadas por los hablantes

[8] Nótese que las EMP con verbo soporte se reproducen como aparecen en el corpus y no han sido lematizadas (en la mayoría de los casos, el infinitivo se destaca). Esta misma observación se aplica a las expresiones fraseológicas verbales que se detallan a continuación.

nativos incluyen: *irse_de_la_casa_de_sus_padres*; *navegar_la_sociedad_de_hoy*; *dieron_las_herramientas*; *aprender_de_los_errores_que_cometemos*; *dar_lo_mejor_de_sí*; *no_todo_es_color_de_rosa*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *caminando_de_puntillas*; *viven_en_el_momento*; *pasar_un_buen_momento*; *ponerse_pesado*; *elegido_caminos_diferentes*; *viene_a_la_mente*; *interponen_en_el_camino*.

Nótese que en algunos casos la EMP es solo parcialmente fija. Este es el caso de muchas EMP SVC en el que el determinante puede variar libremente (*tener_ mucho/poco_cuidado_de_algo*) o de EMP VIDIOM en las que no solo el sujeto, sino también la distribución del complemento es libre en algunas instancias (algo *viene_a_la_mente* de alguien).

En la Tabla 2, se muestra el total de EMP (y los tipos de EMP) identificadas en el corpus con la distribución de frecuencia normalizada por mil palabras.

EMP	HN	Frec. mil	HNN	Frec. mil
ADV	107	18,325	101	13,003
CONJ	1	171	9	1,159
N	102	17,469	191	24,591
NE	33	5,652	33	4,249
PREP	7	1,199	21	2,704
PRON	2	343	13	1,674
SVC	43	7,364	70	9,012
VIDIOM	27	4,624	19	2,446
Total	322	55,146	457	58,839
Número total de palabras	5.839		7.767	

Tabla 2. Distribución y frecuencia normalizada por mil palabras (Frec. mil) de los tipos de expresiones multipalabra (EMP) por categorías y grupo de participantes: hablantes nativos (HN) y hablantes no nativos (HNN)

Un total de 779 EMP fueron identificadas en el corpus de textos escritos por parte de los hablantes nativos y los no nativos (número total de palabras: 13.606). En la tabla presentada arriba se observa que los hablantes nativos contribuyen con el 41 % de las EMP identificadas mientras que los no nativos contribuyen con el 59 %. Según la incidencia de EMP y su distribución de frecuencia normalizada por mil palabras (55,146), los hablantes nativos presentan un mayor uso de adverbios (ADV, 107; 18,325), sustantivos comunes compuestos (N, 102; 17,469), construcciones con verbo soporte (SVC, 43; 7,364) y expresiones fraseológicas verbales (VIDIOM, 27; 4,624). En el corpus de los hablantes no nativos (Frec. mil: 58,839), se observan con mayor incidencia los adverbios (ADV, 101; 13,003), sustantivos comunes compuestos (N, 191; 24,591) seguido por construcciones con verbo (SVC, 70; 9,012).

Dos muestras de textos escritos acompañan la Figura 1, para resaltar algunas de las EMP (enlazadas y debidamente etiquetadas) ya explicadas:

Muestra de ensayo escrito por un Hablante Nativo (HN): Lo que me hace preguntarme, ¿realmente estamos listos para ser padres? O incluso ¿qué es el ser un buen padre? *Desde_mi_perspectiva/ADV*, todo padre de familia ha *dado_lo_mejor_de_sí/VIDIOM* pero aun así han afectado a sus hijos a través de estas creencias heredadas. Un claro ejemplo es la creencia de que el padre no debe llorar o mostrar alguna *señal_de_afecto/N*, causando una herida de abandono o creando dureza en un niño. Es por eso por lo que cuando veo jóvenes parejas que construyen familias desde una edad muytemprana, lo primero que me *viene_a_la_mente/VIDIOM* es "aún no estamos listos, ¿porque queremos correr?"

Muestra de ensayo escrito por un Hablante No Nativo (HNN): *Por_lo_tanto/ADV*, aprovechar las características de los lugares donde vive la gente puede ser una gran manera de evitar la dependencia en la tecnología para conocer gente nueva. Cuando se pueden utilizar otros *medios_de_interacción_social/N*, la dinámica de las interacciones de los pueblos cambia. Las personas pueden comportarse de forma más natural entre sí porque no hay *pantallas_de_computadora/N* ni aplicaciones que se *interponen_en_el_camino/VIDIOM*. Cuando tienes una conversación *por_Internet/ADV* no estás cada vez más *cerca_de/PREP* la otra persona.
Por_eso/ADV, te recomiendo explorar el rascacielos que siempre habías pensado que era hermoso con alguien, y *harás_conexiones/SVC* reales con la gente.

Figura 1. Muestras balanceadas de ensayos escritos por un hablante nativo (HN) y un hablante no nativo (HNN).

Ya presentada la descripción del corpus, la selección de los tipos de EMP y el esquema de anotación seleccionado, discutido y aceptado por ambos lingüistas, se procede a explicar los experimentos realizados con la herramienta de minería de datos Orange.

2.5. Experimentos de aprendizaje automático para la clasificación de textos

Para determinar el nivel de competencia de los hablantes nativos y los no nativos nos centramos en un problema de clasificación con dos niveles. En este trabajo, se adoptó un enfoque en herramientas de aprendizaje automático para investigar el impacto de EMP en el proceso de clasificación, usando como datos la información del corpus descrito en la Sección 2.3.

Dado que la composición de los ensayos de ambos grupos, en cuanto al número de caracteres, varía ampliamente, de 383 a 4.711 caracteres, con una media de 1.987 caracteres y una desviación media de 1.056, según datos presentados en la Tabla 1, la preparación de datos consistió en la división de cada ensayo en varios segmentos de tamaño similar, dejando intactas las frases completas. Los segmentos tienen entre 600 y 700 caracteres, con un promedio de 636 caracteres y un máximo de 985 caracteres.

Para los experimentos de aprendizaje automático, usamos la herramienta de aprendizaje automática Orange (Demšar *et al.* 2013). Este sistema es completo, fácil de usar y contiene los elementos básicos requeridos para el preprocesamiento y gestión de datos, así como también los algoritmos de aprendizaje utilizados más comúnmente, incluyendo uno de redes neuronales (*Neural Network;* NN por sus siglas en inglés). Orange dispone de varias herramientas de visualización de datos que permiten un análisis más detallado en cuanto al impacto de las EMP en la clasificación de textos de estudiantes nativos y no nativos. El flujo de trabajo básico adoptado para cada uno de los experimentos se muestra a continuación en la Figura 2.

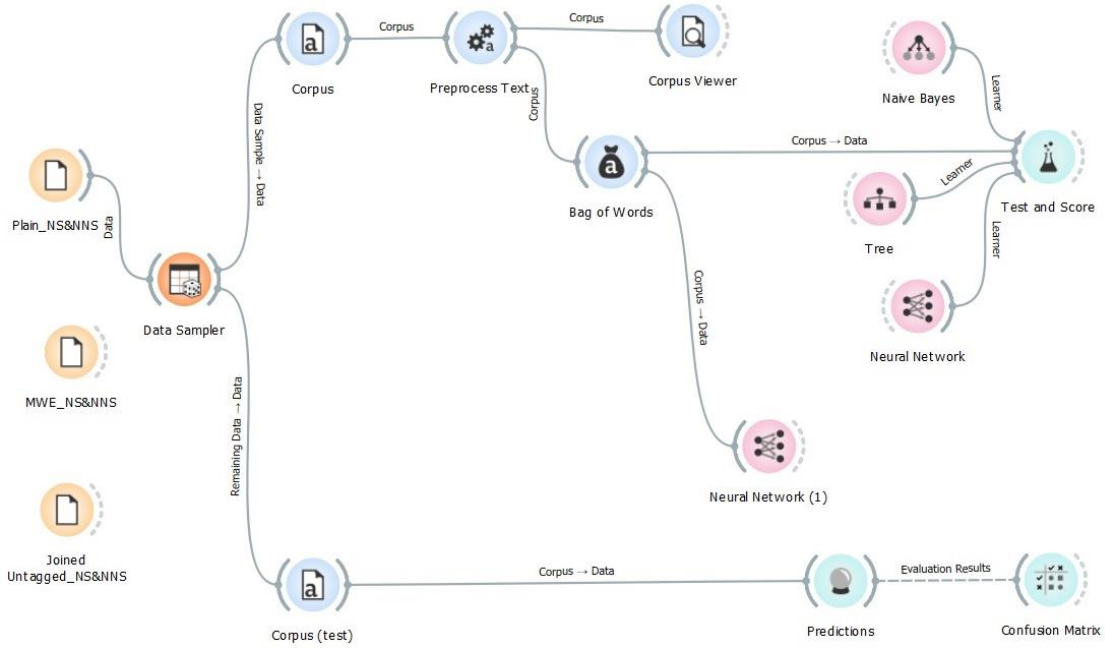


Figura 2. Configuración del flujo de trabajo en Orange para las fases de entrenamiento y prueba del modelo experimental

La Figura 2 muestra (a la izquierda) las tres versiones del corpus utilizadas en este estudio: (i) sin etiquetaje; (ii) con etiquetaje y enlace de las EMP como un solo *token* con el símbolo de barra baja ('_'); (iii) sin etiquetaje, pero con enlace de la EMP y que corresponden a los tres experimentos básicos explicados a continuación. Le sigue un *widget* de muestreo de datos (*Data Sampler*, en inglés, tal como lo muestra la Figura 2) que divide aleatoriamente el corpus en dos particiones: 75 % para la fase de entrenamiento y 25 % para la fase de prueba.

Los tres experimentos efectuados inicialmente con el fin de estimar, de manera general, el efecto de las EMP en el criterio de clasificación por nivel de los hablantes nativos y los no nativos fueron:

- **Experimento 1**, con el corpus de ensayos escritos por los hablantes nativos y los no nativos, sin etiquetaje ni enlace de las EMP.
- **Experimento 2**, con el corpus de ensayos con etiquetaje y enlace de todas las EMP (779). Con base en este diseño, se efectuaron experimentos adicionales que

incluyeron la eliminación, por separado, de los tipos EMP y otros con la eliminación de dos tipos de EMP, con sus respectivos enlaces. Estos experimentos (Exp 4 al Exp 18) se detallan en la Tabla 3.

- **Experimento 3**, con el corpus de ensayos escritos por los hablantes nativos y los no nativos, sin etiquetaje, pero con las EMP enlazadas.

La Figura 2 muestra también el conjunto de modelos de aprendizaje utilizados: *Naïve Bayes* (NB), *Neural Network* (NN), (*Decision*) *Tree*. El *widget* de prueba y puntuación (en inglés, *Test & Score*) se utilizó para entrenar y probar los modelos, solo con los datos del corpus de entrenamiento (partición con el 75 % del corpus), con el fin de seleccionar aquel con el mejor desempeño basado en la precisión de clasificación. Tras varios experimentos, se seleccionó el modelo NN, ya que mostró sistemáticamente mejores resultados. Da Corte & Baptista (2022a) también observaron el mismo desempeño del modelo NN en experimentos similares en un contexto parecido, pero con estudiantes en un curso de desarrollo de gramática en inglés. Se utilizó el *widget* de predicciones (en inglés, *Predictions*) para aplicar al modelo NN los datos de prueba no vistos previamente (partición con el 25 % del corpus) y evaluar la clasificación de los hablantes nativos y los no nativos en un escenario realista. Se utilizó el *widget* matriz de confusión (en inglés, *Confusion Matrix*) para inspeccionar en profundidad los resultados obtenidos.

En el modelo de aprendizaje diseñado para los tres experimentos, utilizando el *widget* de muestreo, los datos se dividieron en tres particiones para ejecutar una validación cruzada, dejando 2/3 para el entrenamiento y 1/3 para las pruebas de predicción de clasificación por nivel. Las tareas de preprocesamiento de datos consistieron en: (i) la *tokenización*, manteniendo tanto las palabras como los signos de puntuación como *tokens*, ya que se estima que la puntuación es un buen indicador del nivel de desarrollo de las habilidades escritas a nivel académico; (ii) la selección del *Averaged Perceptron Tagger* como etiquetador de partes del discurso (*Part-of-Speech Tagge*r; PoS por sus siglas en inglés). No se utilizó ninguna transformación de palabras, ya que la distinción entre mayúsculas y minúsculas puede ser relevante en la clasificación por nivel entre los hablantes nativos y los no nativos en tareas de redacción con fines académicos.

Como paso adicional en la etapa de preprocesamiento, se verificó y comprobó si el filtrado de palabras reservadas (en inglés, *stopwords*) repercutió en los resultados. Para ello, se utilizó un listado estándar de estas palabras en español disponible en línea[9]. En la mayoría de las configuraciones experimentales, la eliminación de las palabras reservadas, tal como se demuestra en la Tabla 3 (Exp 1bis y Exp 3bis), no mejoró ni empeoró los resultados, confirmando así que las palabras reservadas no influyen significativamente en la información semántica del texto en el que aparecen (HaCohen-Kerner *et al*. 2021). Sin embargo, utilizando el corpus etiquetado (Exp 2bis) se alcanzó el mayor nivel de precisión en la fase de entrenamiento (0,892), aunque en la fase de prueba, se observó uno de los resultados más débiles (0,711). Este último nivel de precisión es significativamente inferior al del experimento que

[9] <https://countwordsfree.com/stopwords/spanish>.

retuvo las palabras reservadas (Exp 2: 0,816). Por lo tanto, se decidió omitir el filtrado de las mismas en el resto de los experimentos.

Se compararon también dos métodos de representación de datos que se encuentran disponibles en la caja de herramientas Orange: la bolsa de palabras (*Bag-of-Words* o BoW, por sus siglas en inglés) y la integración de documentos (*Document Embeddings*, o DE, por sus siglas en inglés). En el método BoW, solo se tiene en cuenta la frecuencia de las palabras del texto, ignorando su orden relativo. En el método DE, se tienen en cuenta tanto la frecuencia como la información de co-ocurrencia de las palabras, ya que el texto se convierte en un vector de datos. Se confirmó que el método de la bolsa de palabras produce sistemáticamente mejores resultados. Con resultados similares obtenidos por HaCohen-Kerner *et al*. (2021), se observa, por ejemplo, en la Tabla 3, los resultados del Exp 2tri que corresponden a la utilización del método de representación de datos por medio de integración de documentos y que son significativamente peores que el Exp 2: 0,662 en la fase de entrenamiento y 0,5 en la fase de prueba. Consecuentemente, se descartó la opción de integración de documentos como método de representación de datos y se mantuvo solo la opción de bolsa de palabras (sin palabras reservadas) en el resto de los experimentos.

Corpus (Ensayos segmentados)	**Exp**	**T&S**	**Pred #**	**EMP**	
Sin etiquetaje	**Exp 1**	0,797	**0,842**	0	
Sin etiquetaje (con pr)	Exp 1bis	0,878	**0,816**	0	
Con etiquetaje y unión	**Exp 2**	**0,824**	0,816	779	
Con etiquetaje y unión (sin pr)	Exp 2bis	**0,892**	0,711	779	
Con etiquetaje y unión (con id)	Exp 2tri	0,662	0,500	779	
Sin etiquetaje con unión	**Exp 3**	0,784	**0,842**	779	
Sin etiquetaje con unión (sin pr)	Exp 3bis	0,770	**0,816**	779	%EMP
~ sin SVC	Exp 4	0,824	0,789	90	0,12
~ sin VIDIOM	Exp 5	**0,770**	0,789	41	0,05
~ sin ADV	Exp 6	0,811	0,789	183	0,23
~ sin PREP	Exp 7	0,797	0,789	24	0,03
~ sin CONJ	Exp 8	0,824	**0,763**	7	0,01
~ sin NE	Exp 9	0,838	0,816	52	0,07
~ sin N	Exp 10	**0,770**	0,816	289	0,37
~ sin PRON	Exp 11	0,838	0,816	13	0,02
~ sin SVC&VIDIOM	Exp 12	0,824	0,842	131	0,17
~ sin CONJ&VIDIOM	Exp 13	**0,784**	0,789	48	0,06
~ sin CONJ&SVC	Exp 14	0,851	0,816	97	0,12
~ sin PREP&ADV	Exp 15	0,811	0,763	207	0,27
~ sin PREP&CONJ	Exp 16	0,851	0,816	31	0,04
~ sin ADV&CONJ	Exp 17	0,838	0,737	190	0,24
~ sin N&VIDIOM	Exp 18	**0,784**	0,816	330	0,40

Tabla 3. Resultados obtenidos en las fases de entrenamiento (T&S) y prueba (Pred) en los tres diseños experimentales (Exp 1, Exp 2 y Exp 3) y experimentos complementarios (~ bis, ~ tri) seguidos de los experimentos derivados del Exp 2 y con la eliminación sucesiva (una a la vez) decada tipo EMP y de diferentes combinaciones de tipos de EMP[10]

[10] Leyenda: pr = palabras reservadas, id = integración de documentos.

3. Análisis de los resultados

En los dos primeros experimentos (corpus sin etiquetaje y corpus con etiquetaje y enlace), la adición de las etiquetas de EMP sugiere una mejoría en el nivel de rendimiento del modelo simple establecido como base, ya que pasa, en la fase de entrenamiento, de una precisión de clasificación de 0,797 a 0,824. Sin embargo, en la fase de prueba, la precisión del modelo con el corpus sin etiquetar aumenta ligeramente en comparación al corpus con el etiquetado. Esto puede deberse a una combinación de factores: (i) el hecho de utilizar todas las etiquetas juntas pudo haber degradado la precisión de los resultados; (ii) el pequeño tamaño del corpus y el muestreo aleatorio de la parte del corpus no etiquetada utilizada para las pruebas pudo haber sesgado los resultados. Además, dado que el corpus usado en los experimentos incluía las etiquetas de las EMP, que luego se utilizaron en la bolsa de palabras en la etapa de preprocesamiento como *tokens*, no se descarta que su presencia haya influido en los resultados.

Para entender el impacto que tienen las etiquetas en la clasificación y corroborar las suposiciones mencionadas, se llevó a cabo el Experimento 3, manteniendo como un único *token* la EMP (enlazada) y eliminando todas las etiquetas (8) seleccionadas. En este experimento, el modelo reconoce la EMP presente en el corpus, mas no considera las etiquetas con los tipos de expresiones. Los resultados de este experimento muestran una ligera disminución en el rendimiento del modelo en la fase de aprendizaje (0,784), por debajo de la base, pero coincide con esta en la fase de prueba (0,842).

Como parte de este estudio, se realizaron ocho experimentos adicionales (Experimentos 4-11) con el fin de evaluar el impacto de cada tipo de EMP en los resultados generales descritos anteriormente. Para ello, adoptamos una estrategia de *one-out* o 'una a la vez', eliminando una etiqueta de EMP a la vez, por ejemplo, VIDIOM, ejecutando el modelo de aprendizaje con la misma configuración y manteniendo las otras siete EMP (etiquetadas y enlazadas). De este modo, el modelo reconoce la misma información asociada con el Experimento 2, excepto la información relativa al tipo de EMP que se está poniendo a prueba. Dado que cada tipo de EMP está distribuido de forma diferente en el corpus, presentamos en las columnas de la izquierda, en la Tabla 3, el número y el porcentaje de EMP eliminadas en cada experimento. El argumento para interpretar los resultados es el siguiente: cuanto mayor sea la caída o disminución de la precisión del criterio de clasificación en los resultados generales de cada experimento, mayor es la contribución de ese tipo de EMP (en particular) al rendimiento del modelo.

Estos resultados fueron comparados con el número de EMP que quedaron en el corpus en cada experimento, al utilizar la estrategia *one-out*. Se calculó y encontró un coeficiente de correlación de Pearson débil y positivo de 0,259 entre la precisión de los modelos en la fase de entrenamiento y el número de EMP. Al comparar los resultados de la fase de prueba con el número de EMP en el corpus, se encuentra un coeficiente de Pearson más interesante, pero aun así moderado y negativo de -0,349. Este coeficiente parece indicar que en la fase de entrenamiento, como se esperaba, cuanto mayor es el número de EMP disponibles, mejores son los resultados de precisión que produce el modelo. Sin embargo, probablemente debido a que el muestreo

se realizó con un corpus pequeño, se observa el efecto contrario en la fase de prueba, lo que limita llegar a conclusiones más definitivas al respecto.

Los resultados muestran que la eliminación de los tipos de expresiones fraseológicas verbales (VIDIOM) y sustantivos comunes compuestos (N) producen la mayor disminución en la precisión (0,054) de los modelos en la fase de aprendizaje. Las expresiones VIDIOM son usadas por los hablantes nativos y los no nativos, en cuanto a frecuencia, de manera comparable: 27 incidencias por parte de los nativos y 19 por los no nativos. Esta mínima diferencia alude al nivel similar (y avanzado) de competencia entre los dos grupos. En cuanto al uso de sustantivos compuestos, los no nativos muestran un mayor uso de estas EMP, con 191 incidencias, en comparación con 102 por parte de los nativos.

En cuanto a la precisión de clasificación, le siguen inmediatamente los resultados obtenidos con la eliminación de las preposiciones (PREP) (0,027). Los experimentos con la eliminación de los tipos de EMP construcciones con verbo soporte (SVC) y locuciones conjuntivas (CONJ) no modificaron la precisión (0,824) del modelo base (Experimento 2). Por último, los experimentos con nombres de entidades (NE) y pronombres (PRON) mostraron una diferencia positiva (0,838) con respecto al modelo base, sugiriendo que la presencia de estas expresiones puede dificultar la evaluación de las destrezas escritas de los hablantes nativos y los no nativos. Sin embargo, en la fase de pruebas, es la eliminación de CONJ la que produce la mayor disminución en el rendimiento del modelo (0,053), situándose los tipos de EMP SVC, VIDIOM, ADV y PREP en segundo lugar, ex aequo (0,027) y sin ningún cambio en las categorías restantes NE, N y PRON.

Dado que la combinación de EMP puede tener una mayor incidencia en la precisión de la clasificación que una sola categoría, se realizaron siete experimentos adicionales (Exp 12-18, Tabla 3). Estos experimentos siguen las ideas expuestas por Siyanova-Chanturia & Spina (2020), en un estudio a gran escala de EMP en ensayos escritos por aprendices de una segunda lengua (italiano). Estos autores afirmaron que diferentes combinaciones de patrones de EMP «pueden captar una descripción más compleja y detallada de los patrones de desarrollos de hablantes no nativos» (Siyanova-Chanturia & Spina 2020: 453, traducción propia). Se procedió a adoptar una estrategia de *two-out* o 'dos a la vez': eliminando dos etiquetas de EMP a la vez con sus respectivos enlaces y ejecutando el modelo de aprendizaje con la misma configuración y manteniendo etiquetadas las EMP restantes. Las combinaciones de CONJ&VIDIOM (Exp 13) y N&VIDIOM (Exp 18) fueron las que reportaron la mayor reducción en el nivel de clasificación en la fase de entrenamiento (0,784), siendo N&VIDIOM la que mayor peso tuvo en cuanto al número de EMP (40 %), en relación con el total de EMP en todo el corpus.

4. Conclusiones y futuros estudios

Los tipos de EMP VIDIOM (expresiones fraseológicas verbales) y N (sustantivos comunes compuestos) tienen un mayor impacto en el modelo de aprendizaje en la fase de entrenamiento, mientras que, a pesar del pequeño número de elementos

presentes en el corpus utilizado, el tipo de EMP CONJ (locuciones conjuntivas) parece tener el mayor impacto en la fase de prueba. Cuando se combinaron y eliminaron dos tipos de EMP a la vez, las combinaciones CONJ&VIDIOM y N&VIDIOM fueron las que generaron la mayor reducción en el nivel de clasificación en la fase de entrenamiento, siendo N&VIDIOM la combinación que mayor peso tuvo en cuanto al número de EMP presentes en todo el corpus.

Estos resultados, aunque todavía provisionales y extraídos de un corpus pequeño, pueden indicar áreas de desarrollo léxico (y tipos de EMP) que deberían tomarse en cuenta para enriquecer las estrategias pedagógicas y de competencia lingüística en los programas de estudios orientados a la enseñanza del español como lengua extranjera. Aunque muchos estudios se ocupan de esta temática, estos suelen centrarse en las destrezas de lectoescritura en una segunda lengua. En este estudio también destacamos la importancia de estas habilidades para quienes tienen el español como lengua materna.

En el contexto de Oklahoma (4.019.800 habitantes),[11] existe una gran demanda de intérpretes y traductores calificados para desempeñar ambas tareas en español e inglés, particularmente, en el ámbito médico y legal (en este artículo no abordamos terminología técnica). Esta demanda se debe al aumento de hispanohablantes (actualmente, 11,8 % de la población) que tienen un nivel bajo de competencia en inglés (36,6 %)[12]. Por consiguiente, es necesario constatar que los participantes en los programas de certificación de traducción e interpretación: primero, estén en el nivel correcto y reciban la formación adecuada en cuanto al español como lengua extranjera; y, segundo, puedan comunicarse debidamente en dicha lengua como lengua meta o de destino (en inglés, *target language*).

En este análisis se tomó en cuenta un curso de desarrollo enfocado en la gramática y composición escrita de textos en español. Nos proponemos realizar un estudio de cursos de desarrollo en otros idiomas, como, por ejemplo, el inglés, para investigar si la eliminación o combinación de tipos de EMP arrojan resultados más significativos en cuanto a la clasificación del nivel de competencia lingüística de los participantes. Adicionalmente, sería de interés identificar si la disminución o aumento del uso de EMP específicas por parte de los hablantes nativos versus no nativos se debe al reconocimiento de dichas palabras o si su uso se ve limitado por los temas sugeridos para la elaboración de ensayos.

Referencias bibliográficas

ACTFL (2016), *Assigning CEFR ratings to ACTFL assessments* [disponible en <https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Ass essments.pdf>, 7/3/2023].

Alexopoulou, Theodora – Michel, Marije – Murakami, Akira – Meurers, Detmar (2017), «Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques», *Language Learning* 67(S1), 180-208.

[11] <https://www.census.gov/quickfacts/OK>.

[12] <https://www.migrationpolicy.org/data/state-profiles/state/language/OK>.

Alpine Testing Solutions (2020), *Examination of the ACTFL Writing Proficiency Test (WPT) in English, Russian, and Spanish for the ACE Review – Part B: Statistical Analysis & Evidence of Validity*, Orem, UT: Alpine Testing Solutions.

Arnold, Taylor – Ballier, Nicolas – Gaillat, Thomas – Lissòn, Paula (2018), «Predicting CEFRL levels in learner English on the basis of metrics and full texts», *Proceedings of the 20th Conférence Sur l'Apprentissage Automatique*. INSA de Rouen, 20-22 June 2018, *ArXiv:1806.11099*.

Baptista, Jorge – Mamede, Nuno – Reis, Sonia (2022), «Support Verb Constructions across the Ocean Sea», *Proceedings of the 18th Workshop on Multiword Expressions @* LREC2022, Marseille, France. European Language Resources Association, 26-36.

Consejo de Europa (2002/2020), *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación,* [disponible en <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf>; <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco_complementario/mcer_volumen-complementario.pdf>, 7/3/2023].

Corpas Pastor, Gloria (2017), «Collocational constructions in translated Spanish: what corpora reveal», en Mitkov, R. (ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017,* Londres: Springer, 29-40.

Da Corte, Miguel – Baptista, Jorge (2022a), «A Phraseology Approach in Developmental Education Placement», en Corpas Pastor, G. – Mitkov, R. – Kunilovskaya, M. – Caro Quintana, R. (eds.) *Computational and Corpus-based Phraseology*, Proceedings of EUROPHRAS 2022, Malaga, September 28-30, 2022, Londres: Springer, 79-86.

Da Corte, Miguel – Baptista, Jorge (2022b), «Lista de expresiones multipalabra detectadas en ensayos escritos en un curso de desarrollo de gramática y composición» [disponible en <https://doi.org/10.13140/RG.2.2.35177.57443>, 7/3/2023].

Dahunsi, Toyese Najeem – Ewata, Thompson Olusegun (2022), «An exploration of the structural and colligational characteristics of lexical bundles in L1–L2 corpora for English language teaching», *Language Teaching Research* 1-17 [disponible en <https://doi.org/10.1177/13621688211066572>, 7/3/2023].

Demšar, Janez – Curk, Tomaž – Erjavec, Aleš – Gorup, Črt – Hočevar, Tomaž – Milutinovič, Mitar – Možina, Martin – Polajnar, Matija – Toplak, Marko – Starič, Anže – Stajdohar, Miha – Umek, Lan – Žagar, Lan – Žbontar, Jure – Žitnik, Marinka – Zupan, Blaž (2013), «Orange: data mining toolbox in Python», *The Journal of machine Learning research* 14(1), 2349-2353.

El-Dakhs, Dina Abel Salam – Khan, Shazia Khalid – Al-Khodair, Maram (2022), «Do foreign language learners mine input texts for multiword expressions? The case of writing story retellings», *Ampersand* 9, 100080.

Erdmann, Alexander – Wrisley, David Joseph – Brown, Christopher – Cohen-Bodénès, Sophie – Elsner, Micha – Feng, Yukun – Brian, Joseph – Joyeux-Prunel, Béatrice – de Marneffe, Marie-Catherine (2019), «Practical, efficient, and customizable active learning for named entity recognition in the digital humanities», *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, 2223-2234.

Esfandiari, Rajab – Ahmadi, Mohammad (2022), «Phraseological Complexity and Academic Writing Proficiency in Abstracts Authored by Student and Expert Writers», *English Teaching & Learning*, 1-20.

Farahmand, Meghdad – Smith, Aaron – Nivre, Joakim (2015), «A multiword expression data set: annotating non-compositionality and conventionalization for English noun compounds», *Proceedings of the 11th Workshop on Multiword Expressions,* 29-33.

Fotopoulou, Aggeliki – Laporte, Éric – Nakamura, Takuya (2021), «Where Do Aspectual Variants of Light Verb Constructions Belong?», *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021),* 2-12.

García-Page, Mario (2008), *Introducción a la fraseología española: estudio de las locuciones*, Barcelona: Anthropos.

Gross, Maurice (1996), «Lexicon-grammar», en Brown, K. – Miller, J. (eds.), *Concise Encyclopedia of Syntactic Theories*, Cambridge: Pergamon, 244-259.

HaCohen-Kerner, Yaakov – Miller, Daniel – Yigal, Yair (2020), «The influence of preprocessing on text classification using a bag-of-words representation». *PloS ONE* 15(5), e0232525 [disponible en <https://doi.org/10.1371/journal.pone.0232525>, 7/3/2023].

Hernández, Mireia – Costa, Alber – Arnon, Inbal (2016), «More than words: multiword frequency effects in non-native speakers», *Language, Cognition and Neuroscience* 31(6), 785-800.

Hinkel, Eli (2023). «Teaching and Learning Multiword Expressions», *Handbook of Practical Second Language Teaching and Learning,* Nueva York: Routledge, 435-448.

Kochmar, Ekaterina – Gooding, Sian – Shardlow, Matthew (2020), «Detecting multiword expression type helps lexical complexity assessment», *LREC 2020: Proceedings of the 12th Conference on Language Resources and Evaluation,* The European Language Resources Association (ELRA), 4426-4435.

Laporte, Éric (2018), *Choosing features for classifying multiword expressions,* en Sailer, M. – Markantonatou, S. (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, Berlín: Language Science Press, 143-186.

Nam, Daeyeon – Park, Kwanghyun (2020), «I will write about: Investigating multiword expressions in prospective students' argumentative writing», *Plos one* 15(12), e0242843.

Pasquer, Caroline – Savary, Agata – Ramisch, Carlos – Antoine, Jean-Yves (2020), «Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?», *Proceedings of the 28th International Conference on Computational Linguistics,* 3333-3345.

Savary, Agata – Cordeiro, Silvio Ricardo – Ramisch, Carlos (2019), «Without lexicons, multiword expression identification will never fly: A position statement», *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019).* Association for Computational Linguistics, 79-91.

Siyanova-Chanturia, Anna – Spina, Stefania (2020), «Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study», *Language Learning* 70(2), 420-463.

Wilkens, Rodrigo – Seibert, Daiane – Wang, Xiaoou – François, Thomas (2022), «MWE for Essay Scoring English as a Foreign Language», *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference,* 62-69.

Section 2
Advancing Writing Proficiency Classification with NLP-extracted Features and ML Models

Building on the proposed DevEd (annotation) framework introduced in Section 1, which centered on identifying linguistic features characteristic of developing writers, alongside indicators of proficiency such as MWE, *Section 2*, operationalizes and expands this framework by incorporating a broader set of linguistic features relevant to academic writing. These include syntactic complexity, lexical sophistication, and discourse cohesion, which are automatically extracted using Coh-Metrix and CTAP, two widely used Natural Language Processing (NLP) tools for analyzing cohesion and syntactic development. Combined, the DevEd-specific features and those sourced from Coh-Metrix and CTAP offer a more robust linguistically grounded approach to evaluate student writing in automated placement contexts.

The goal of this section is to examine whether this enriched feature set, when used to train and test Machine Learning (ML) models, cannot only replicate but also improve the classification precision of Accuplacer. Specifically, the aim is to evaluate whether these linguistically grounded features can more precisely identify students whose writing demonstrates college readiness, distinguishing them from those who may require additional instructional support, and place them in the appropriate courses. This section presents findings from three studies and addresses the second research question (RQ2): *To what extent can linguistically grounded features, both human-annotated and automatically extracted (using NLP tools), improve placement accuracy when modeled using ML*?

The first study (Da Corte and Baptista (2024b)), titled *Enhancing Writing Proficiency Classification in Developmental Education: the Quest for Accuracy*, sets the stage by first assessing how closely Accuplacer's automated classifications align with expert human judgment. To this end, Accuplacer's essay placements were compared to those produced by six trained human raters across the 100 student writing samples forming the original corpus mentioned in (Da Corte and Baptista (2024a)). Both assessments, automated and human, relied on the same six broad linguistic descriptors currently used by Accuplacer (The College Board, 2022, p. 24), which include: (i) *mechanical conventions*; (ii) *sentence variety & style*; (iii) *development & support*; (iv) *organization & structure*; (v) *purpose & focus*; and (vi) *critical thinking*. The human rating task involved two main steps: (1) evaluating each essay using the six criteria, and (2) assigning an overall placement classification: DevEd Level 1, DevEd Level 2, or College Level. To manage workload and ensure assessment consistency, essays were randomly assigned to raters. Of the 100 essays, 49 were independently evaluated by at least two raters, resulting in five rater pairs.

While ACCUPLACER's descriptors cover broad dimensions of writing, its criteria lack explicit (and measurable) linguistic features, as well as clarity on whether (and how) these features are weighted in the automated scoring process. To address this lack of transparency, two trained linguists created illustrative examples, grounded in authentic student writing and absent from ACCUPLACER's official documentation, to clarify and enhance select descriptors[35]. These examples, together with detailed annotation protocols, were compiled into a publicly available guidelines document[36] (Da Corte and Baptista (2024b)) to promote standardized interpretation and consistent human scoring. Before commencing the annotation task, the raters, who were familiar with DevEd and ACCUPLACER placement practices in higher education, completed comprehensive training led by one of the guidelines' authors. This training covered task expectations, ethical considerations, the use of the guidelines, and the overall annotation procedures.

With the guidelines established and training sessions completed, all essays were assessed. Intraclass Correlation Coefficients (ICC) were calculated for the 49 essays to gain insights into the agreement between raters within each pair, for each linguistic descriptor[37]. Low ICC scores were expected, especially for more complex (and abstract) linguistic features like *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 (and were enhanced with examples). An ICC calculation was also performed to determine how human classification (all individual ratings) correlated with that of ACCUPLACER. An ICC score of 0.119 was obtained, indicating a *poor*[38] correlation between humans and this automated system in the assessment of all six linguistic features.

Despite the enhancements made to complex criteria such as *mechanical conventions* and *organization & structure*, persistent rating challenges were observed, and among the 49 essays assessed by at least two raters, only 27 (55%) received the same overall placement classification (DevEd Level 1 or DevEd Level 2), revealing ongoing inconsistency in human judgments. Pearson correlation coefficients further confirmed *weak* alignment between human and automated scores (Pearson r=0.172). These findings suggest that even when using enhanced guidelines and the same six criteria ACCUPLACER, humans struggle to produce consistent and reliable placement outcomes. Furthermore, these findings (i) raise questions about the accuracy of ACCUPLACER in reflecting expert human judgment, and (ii) highlight the need for more linguistically grounded and transparent approaches to enhance placement accuracy in DevEd.

The rating task also produced a refined dataset that was later leveraged to train and test four ML algorithms using the data-mining tool ORANGE: Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Neural Network (NN). The goal of these ML experiments was to

[35]These experiments preceded the publication of the 2022 ACCUPLACER Program Manual (The College Board (2022)), which expands on the 2018 edition (The College Board (2018)), without affecting the methods or results reported.

[36]As a reference, these guidelines have also been included in **Appendix B.1**

[37]The details about all ICC calculations, per pair, per linguistic descriptor, are included in the paper, (Da Corte and Baptista (2024b)).

[38]ICC scores were interpreted following the coefficient thresholds set forth in (Koo and Li (2016)).

identify the linguistic descriptors that were more predictive of placement in DevEd. For these experiments, only the 27 essays (55%) on which both raters (per pair) agreed on the overall classification level were included in this dataset. Feature ranking using the Information Gain scoring method identified *idea development & support*, *critical thinking*, and *organization & structure* as the most informative descriptors. These were followed by *sentence variety & style*, *purpose & focus*, and lastly, *mechanical conventions*, which was the least informative feature. Using the top-ranked features, RF achieved the highest classification accuracy (CA) at 88.9%, in the testing phase, indicating that, under these conditions, nearly nine out of ten students would be placed into the appropriate DevEd course. Although RF achieved a high CA score, the results of this study reaffirm the need for greater precision in descriptor definitions and the expansion of classification guidelines.

The previous study identified key limitations, particularly the lack of linguistic detail in ACCUPLACER's descriptors and the inconsistencies in how they were applied by human raters, even when clarified criteria were provided. Now, the focus shifts to the potential of computational tools to offer more consistent and linguistically grounded placement outcomes. When expert human judgment lacks consistency and automated systems lack interpretability, it becomes essential to explore scalable alternatives that rely on transparent, linguistically grounded features. For this reason, the next paper (Da Corte and Baptista (2024c)), titled *Leveraging NLP and Machine Learning for English (L1) Writing Assessment in Developmental Education*, explores NLP tools and ML techniques to identify optimal predictors and linguistic feature combinations that can enhance placement accuracy (Santos et al. (2021)). By improving how students' writing is evaluated and how they are placed as a result of this assessment, fairer opportunities to better support linguistically underprepared students can be afforded, thus reducing disparities in access to college-level coursework (Beaulac and Rosenthal (2019); Qian et al. (2020)).

This study builds on extensive research that leverages ML and NLP techniques to enhance writing assessment and classification (Pal and Pal (2013); Crossley et al. (2017); Filighera et al. (2019); Crossley (2020); Bujang et al. (2021)). Using the same 100-essay corpus employed throughout this research, each essay was parsed through COH-METRIX and CTAP, producing a dataset consisting of a total of 436 linguistic features per essay (106 from COH-METRIX and 330 from CTAP). These were supplemented by 21 DevEd-specific (DES) features, previously introduced in Section 1, including novel indicators such as *Fictional You*, *Fictional We* (Da Corte and Baptista (2024a)), and Multiword Expressions (MWE) (Da Corte and Baptista (2022)).

Multiple supervised ML experiments were conducted to determine which of these linguistic features, and which algorithms, most accurately predict writing placement levels. Feature selection and classification modeling were implemented using the ORANGE data-mining tool, with ten commonly used algorithms being tested. These included Gradient Boosting (GB), Naïve Bayes (NB), Neural Network (NN), and Random Forest (RF). Two classification scenarios[39] were

[39]For consistency and clarity, all *scenarios* and *prompting strategies* in this final thesis are identified alphabetically, e.g., (a), (b), (c), . . .

explored: (a) one scenario using full-length essays (F); and (b) another scenario using the same essays but splitting (S) them into 100-word segments. This second scenario (b) was explored to assess whether text length influenced classification performance or feature ranking. Essays were resampled to maintain balance across levels.

Within each scenario, four experiments[40] were conducted: (1) a baseline Experiment 1, using all the features (457)[41]; (2) a ranked-feature Experiment 2, with the top 11 most discriminative features; (3) a one-out approach Experiment 3, in which features were clustered by type (from within their respective source), and each cluster was sequentially removed; and (4) a final Experiment 4, combining features from Coh-Metrix, CTAP, and DES, using Information Gain to rank and incrementally test the top features in batches of ten until the algorithms reached asymptotic results.

In Experiment 1, a baseline classification accuracy (CA) of 72.7% was established using full (F) samples with RF, using the complete set of features from CTAP. While this is a high CA score, showcasing RF's effectiveness in placement tasks based on writing features (Huang (2023)), it also corresponds to a critical issue: nearly three in ten students remain misclassified in DevEd placement.

In Experiment 2, the top 11 most discriminative features from each source were ranked using the Information Gain scoring method. Texts were then classified using only these top features. Notable gains in classification accuracy (CA) were observed, particularly with the NB and GB models. In the F scenario, both models achieved improvements of nearly 14% over the baseline (72.7%) when using top-ranked features from the Coh-Metrix and CTAP sets. With DES features, the gains were smaller but still positive, reaching 6.1%. In the S scenario, NB showed a 16% accuracy increase with CTAP and nearly 10% with Coh-Metrix, while GB exhibited a more modest gain of almost 5% with DES features. Overall, the results suggest that targeted feature selection enhances classification performance. To better understand their predictive value, in Experiment 3, features were clustered by *type* within their respective platforms.

In Experiment 3, features were grouped into clusters based on linguistic type rather than source. For Coh-Metrix, features were grouped into ten clusters (e.g., Syntactic Complexity, Referential Cohesion, Latent Semantic Analysis). Similarly, CTAP features were grouped into ten clusters (e.g., Lexical Richness, Lexical Sophistication, Syntactic Complexity). The 21 DES features were grouped into four main clusters: Orthographic, Grammatical, Lexical and Semantic, and Discursive. A one-out approach was used to evaluate the impact of each cluster by removing one at a time during classification tasks. To interpret the results, the following principle was followed: the larger the decrease in accuracy, the greater the significance of the removed feature cluster.

[40]The full dataset results of this in-depth analysis are publicly available for transparency and reproducibility in evaluating model performance through GitHub.

[41]A detailed description of the linguistic features extracted from Coh-Metrix and CTAP can be found in the official documentation of each tool.

The impact of holding out these clusters varied across learning models. In the F scenario, CA performance decreases were observed with algorithms tied to lexical and syntactic features. For instance, k-Nearest Neighbors (kNN)'s CA decreased by nearly 23% when the Descriptive (DESC) cluster (including features such as *number of tokens*) was removed. Similarly, Logistic Regression (LR)'s CA decreased by 11% without the Word Information (WORDINFO) cluster, and CN2 Rule Induction (CN2)'s CA decreased by 9.1% with the exclusion of the Syntactic Complexity (SYNTCOMPLEX) cluster.

In contrast, the S scenario demonstrated performance improvements for several models when specific clusters, particularly syntactic and discursive, were excluded. For example, the NN model improved its performance by an average of 13.7%, with the most significant gain (+18.2%) occurring when the Syntactic Pattern Density (SYNTPATTERNDENS) cluster was removed. The GB model improved by 11% on average, with a 12.8% gain attributed to removing SYNTPATTERNDENS. Finally, RF showed a significant improvement in CA (+18.2%) with the removal of the Situational Model (SITMODEL) cluster. These results suggest that in reduced feature scenarios, specific syntactic and discursive clusters may hinder classification accuracy (when they are kept) rather than enhance it.

The last experiment, Experiment 4, consisted of classifying the text sample units by aggregating features from the three distinct sets (Coh-Metrix, CTAP, and DES) and subsequently identifying the most discriminative ones using the Information Gain ranking method. Among the top 15 most predictive features were two DevEd-specific (DES) features—*verb agreement* (VAGREE) and *multiword expressions* (MWE), as well as a selection from CTAP and Coh-Metrix, including *sophisticated noun type ratio (NGSL)*, *prepositional phrases per sentence*, and *particle PoS density* (CTAP); and *CELEX word frequency for content words*, *hypernymy for nouns*, and *text easability PC narrativity percentile* (Coh-Metrix).

Several algorithms, namely the CN2, Decision Tree (DT), GB, LR, and Support Vector Machine (SVM), underperformed relative to the baseline. Among the algorithms, NB stood out as the fastest learning model and consistently performed the best throughout the experiment. With 30 features, the model achieved a CA of 77.3%, which represented an almost 5% enhancement over the baseline (72.7%). The model continued performing above the baseline as more features were added, reaching a peak CA score of 81.8% at 60 features. When the experiment was conducted within the S scenario, only one model, the NN, performed above the baseline (64%) with a CA of 65.6% (with ten features), which is only a 1.6% improvement. As more features were added, scores deteriorated substantially.

These top-predictive features reflect the interconnectedness of lexical precision, syntactic complexity, and discourse coherence, which are core dimensions of academic writing proficiency in the DevEd context. The improvement in classification accuracy achieved in Experiment 4 suggests that, with the selected feature combination, NB could accurately place approximately eight out of ten students. This represents a notable gain compared to the seven out of ten students with the baseline CA score of 72.7%, and marks an attempt (through feature selection) to support

a placement process more closely aligned with students' actual writing proficiency. While technically, two students may still face misclassification, the findings shed light on linguistic features that are more contextually relevant to DevEd and, therefore, support pedagogical practices that can be more responsive to the needs of developing writers.

Having examined the degree of alignment (or lack thereof) between automated and human classifications, shown in the first study; and the predictive value of linguistically grounded features through ML modeling, demonstrated in the second; the final study in this section, titled *Refining English Writing Proficiency Assessment and Placement in Developmental Education Using NLP Tools and Machine Learning* (Da Corte and Baptista (2025c)), expands both the dataset and methodological scope of this investigation. Thus, this study aims to: (i) enhance writing assessment and placement by analyzing an extended corpus of essays using automated and manually annotated linguistic features employed in previous experiments; (ii) evaluate the impact of these combined features on classification accuracy through several ML models; (iii) compare these classification accuracy results against the previous baseline of 72.7% accuracy achieved by RF (Da Corte and Baptista (2024c)); and (iv) use these results to provide actionable insights to better support student placement and instruction in DevEd courses.

While still centered on evaluating the accuracy of automated classification systems, more pecisely ACCUPLACER, this study continues to tackle the persistent uncertainty surrounding how such systems define, weigh, and rank linguistic features (Roscoe et al. (2014); Johnson et al. (2022)). This lack of transparency makes it difficult for students to understand which specific aspects of their writing influenced their placement, what features define college-level writing, and which concrete patterns they should focus on to improve their writing skills. Without this level of clarity and specificity, students are left with limited actionable guidance to develop the skills needed for academic success (Ma et al. (2024)).

The original corpus of 100 student essays used in prior studies was expanded by adding 200 randomly selected texts (100 from DevEd Level 1 and 100 from DevEd Level 2) from a larger pool of 1,000, resulting in a final corpus of 300 essays totaling 97,339 tokens. Balanced with 150 texts per DevEd level, the corpus continues to focus exclusively on non-college-level classifications to improve placement precision for students who need DevEd support, which remains the central focus of this thesis. These additional essays were annotated by the same raters using the same protocols, guidelines, and training established in earlier studies. Each text was manually tagged with the DevEd-specific (DES) features[42], assigned a proficiency classification (DevEd Level 1, DevEd Level 2, or College Level), and parsed through COH-METRIX and CTAP to extract the same automated linguistic features.

In alignment with this study's aim of refining placement accuracy, the original list of 21 DES features was narrowed to 11 using the Information Gain ranking method in ORANGE. These included: grapheme-level errors (ORT), word splits (WORDSPLIT), punctuation and contractions

[42] As an illustration, **Appendix D** presents two annotated sample essay responses (one from DevEd Level 1 and one from DevEd Level 2) from the 200-text corpus expansion.

(PUNCT), word omissions (WORDOMIT), word repetition (WORDREPT), verb agreement (VAGREE), referential pronoun shifts (ALTERN), lexical precision (PRECISION), multiword expressions (MWE), argumentation with reasoning (REASON), and with examples (EXAMPLE). To assess consistency and diagnostic value, the distribution of DES features was examined across the original 100-text corpus, the added 200, and the whole set of 300. Grapheme-level errors (ORT) and argumentation with example (EXAMPLE) consistently ranked among the top three across all subsets, reinforcing their relevance to DevEd placement. Comparisons between human and ACCUPLACER classifications revealed persistently weak correlations across datasets. While ACCUPLACER matched human placement in 79% of the 300 texts, the system struggled to distinguish between DevEd Level 1 and DevEd Level 2. It misclassified 56 texts and failed to identify any College-level texts.

The next phase of analysis employed supervised ML techniques through the ORANGE data-mining toolkit. A total of ten ML algorithms were tested, using the expanded corpus of 300 essays and a refined feature set of 445 linguistic features (106 from COH-METRIX, 328[43] from CTAP, and 11 DES) to: (i) identify the most salient linguistic features for DevEd placement classification; and (ii) determine which ML algorithm achieved the highest classification accuracy (CA). Experiments were conducted under two scenarios: (a) using the full corpus with classification levels assigned by ACCUPLACER, and (b) using the same texts but with the human raters' classification. Each scenario involved initial testing with all 445 features, followed by experiments adding features incrementally and in sets of ten, based on the Information Gain scoring method rankings, until performance plateaued.

A baseline CA of 72.7%, previously achieved with the RF algorithm (Da Corte and Baptista (2024c)), was used as the reference point. In Scenario 1, text samples were classified using ACCUPLACER-assigned levels and the 445-feature set. Seven out of ten ML models surpassed the original 72.7% baseline, with GB reaching 80%, indicating that the combined features hold significant potential for improving student placement. The most predictive features came predominantly from CTAP. To identify which features contributed most to improved placement accuracy, the most discriminative features from COH-METRIX, CTAP, and DES were identified first using the Information Gain ranking method. The top-ranked features, overall, were mostly from CTAP (99%), followed by two COH-METRIX features, and one DES feature, VAGREE ranking 117th. Notably, VAGREE had previously ranked 7th when the experiments were conducted with a smaller corpus size of 100 texts (Da Corte and Baptista (2024c)).

All features were grouped and tested incrementally in bundles of ten. The NN model quickly peaked in the first run, achieving a CA score of 80% with only ten combined features. Testing the feature set incrementally confirmed asymptotic gains. The NB model demonstrated strong performance with consistent CA scores of 77% or higher, achieving asymptotic results after

[43] At the time of the experiments, the following two CTAP features were not available on the platform and, therefore, not incorporated into the analysis: **Number of POS Feature: Punctuation Tokens** and **Number of POS Feature: Punctuation Types**.

a total of 30 combined features. To better gauge the impact of DES features on the overall classification task, the 30 CTAP features, where NB reached its asymptote, were combined with all 11 DES features. Most models improved their performance with the addition of DES features. DT and Adaptive Boosting (AdaBoost) showed the most significant gains with the addition of all DES features (41 features total) (+8.5% and +6.5%), while the NB remained stable with a modest 0.5% gain above the 72.7% baseline.

In Scenario 2, a total of 272 texts (balanced by level; 136 texts from DevEd Level 1 and 136 texts from DevEd Level 2) were classified using all 445 features, with the human-assigned skill levels as the target variable. The number of texts employed here differs from that in Scenario 1, given that human classifications and those of ACCUPLACER were not the same. Overall, CA scores were lower than in Scenario 1, with most models failing to surpass the 72.7% baseline. The SVM model was the highest-performing model, with a score of 71.3%. Most of the highest-ranked features were derived from CTAP, with a much smaller proportion from COH-METRIX. Two DES features, EXAMPLE and ORT, ranked 9th and 31st, which highlight the importance of including human-derived features in the task.

Similar to Scenario 1, features were grouped and tested incrementally in bundles of ten, ranked by their Information Gain scores. Four algorithms (CN2, DT, AdaBoost, and k-Nearest Neighbors [kNN]) underperformed relative to the 72.7% baseline, while others, most notably NB and LR, surpassed it. NB reached nearly 80% accuracy with just ten features, though its performance declined slightly as additional features were added, remaining above the baseline up to the first 30. LR peaked at 78.5% with 40 features, marking an improvement of almost 6% over the baseline. After this experiment was completed, the top 40 features and the remaining nine DES features (two were already included in the top 40) were combined and tested to evaluate whether the addition of more DES features improved the models' performance accuracy. Four ML algorithms (DT, GB, LR, and NN) exhibited declines in CA scores with this feature combination, including LR. Conversely, models such as AdaBoost, RF, kNN, SVM, and NN showed moderate improvements, but their scores remained below the 78.5% threshold. Notably, the NB model matched the highest CA score of 78.5% achieved by LR with 40 features, consistent with its performance from the first scenario.

While some models faced declines, including DES features in DevEd classification tasks continues to add valuable insights into writing performance. It is important to note that the top features in Scenario 2 diverged from those in the ACCUPLACER-based scenario. Here, a larger share came from COH-METRIX (23%) and DevEd-specific features (5%), suggesting that while classification accuracy did not improve as much with human-assigned levels, human raters are more attuned to aspects of writing that involve cohesion, lexical precision, and semantic clarity. Notably, DES features such as EXAMPLE and ORT ranked among the top 31, reinforcing their value in human classification as markers that trained evaluators can more easily identify in student writing.

References to the papers in order of presentation:

Paper 4:
Da Corte, M. and Baptista, J. (2024). **Enhancing Writing Proficiency Classification in Developmental Education: The Quest for Accuracy**. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6134–6143, Torino, Italia. ELRA and ICCL.

Paper 5:
Da Corte, M. and Baptista, J. (2024). **Leveraging NLP and Machine Learning for English (L1) Writing Assessment in Developmental Education**. In Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France, volume 2, pages 128–140.

Acknowledgment of Citation: (Da Corte and Baptista (2024c)) has been cited in a recent publication: *Optimizing Automated Essay Scoring: A Comparative Study of Machine Learning Approaches with a Focus on Ensemble Methods*, published in World Journal of English Language, April 2025, 15(5), p.272. DOI: 10.5430/wjel.v15n5p272, under a CC BY 4.0 license. Authors: Kornwipa Poonpon and Wirapong Chansanam.

Paper 6:
Da Corte, M. and Baptista, J. (2025). **Refining English Writing Proficiency Assessment and Placement in Developmental Education using NLP Tools and Machine Learning**. In Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU, pages 288–303. INSTICC, SciTePress.

Acknowledgment of Citation: (Da Corte and Baptista (2025c)) has been cited in a recent publication: *Explainable Educational Assistant Integrated in Moodle: Automated Semantic Assessment and Adaptive Tutoring Based on NLP and XAI*, published in Discover Artificial Intelligence, July 2025, 5(1). DOI: 10.1007/s44163-025-00438-y, under CC BY-NC-ND 4.0 license. Authors: William Villegas, Rommel Gutierrez, Joselin García-Ortiz, and Vanessa Guevara.

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/

66
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 rat-

ings) 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, Inter-rater 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.

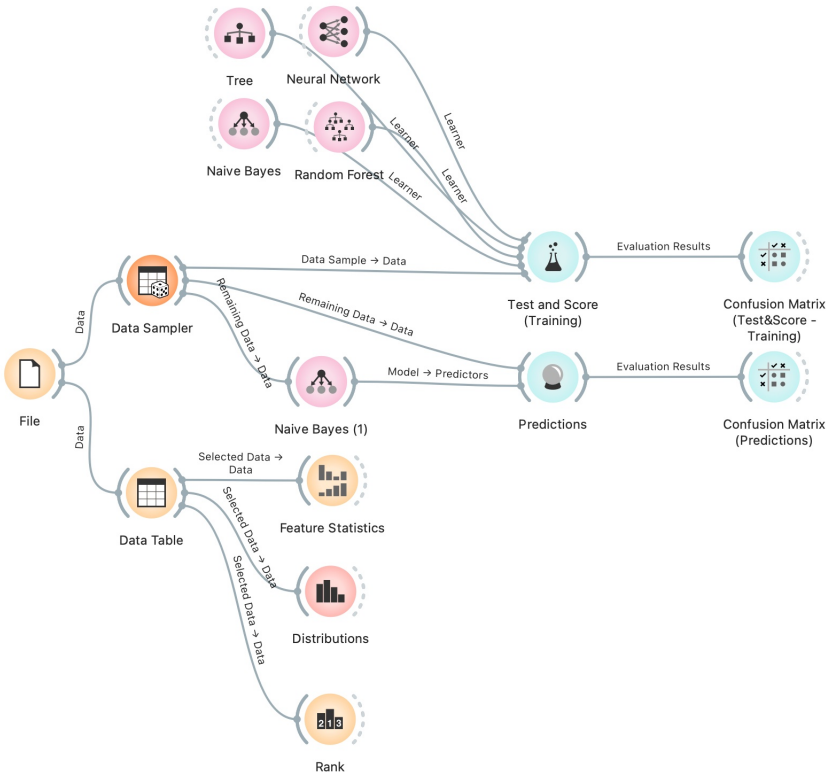


Figure 1: ORANGE Workflow Configuration for Model Training and Testing.

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	χ^2
1	Organization		0.337	11.368
2	Development		0.421	10.726
3	Critical Thinking		0.351	8.425
4	Purpose & Focus		0.196	6.565
5	Sentence Variety		0.199	6.432
6	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.

[9] http://141.225.61.35/CohMetrix2017/

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

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

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.

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.

Leveraging NLP and Machine Learning for English (L1) Writing Assessment in Developmental Education

Miguel Da Corte[1,2][a] and Jorge Baptista[1,2][b]
[1]*University of Algarve, Faro, Portugal*
[2]*INESC-ID Lisboa, Lisbon, Portugal*

Keywords: Developmental Education (DevEd), Automatic Writing Assessment Systems, Natural Language Processing (NLP), Machine-Learning Models.

Abstract: This study investigates using machine learning and linguistic features to predict placements in Developmental Education (DevEd) courses based on English (L1) writing proficiency. Placement in these courses is often performed using systems like ACCUPLACER, which automatically assesses and scores standardized writing assignments in entrance exams. Literature on ACCUPLACER's assessment methods and the features accounted for in the scoring process is scarce. To identify the linguistic features important for placement decisions, 100 essays were randomly selected and analyzed from a pool of essays written by 290 native speakers. A total of 457 Linguistic attributes were extracted using COH-METRIX (106), the Common Text Analysis Platform (CTAP) (330), plus 21 DevEd-specific features produced by the manual annotation of the corpus. Using the ORANGE Text Mining toolkit, several supervised Machine-learning (ML) experiments with two classification scenarios (full and split sample essays) were conducted to determine the best linguistic features and best-performing ML algorithm. Results revealed that the Naive Bayes, with a selection of the 30 highest-ranking features (21 CTAP, 7 COH-METRIX, 2 DevEd-specific) based on the Information Gain scoring method, achieved a classification accuracy (CA) of 77.3%, improving to 81.8% with 60 features. This approach surpassed the baseline accuracy of 72.7% for the full essay scenario, demonstrating enhanced placement accuracy and providing new insights into students' linguistic skills in DevEd.

1 INTRODUCTION AND OBJECTIVES

Developmental Education (DevEd) course models have been implemented in higher education institutions in the United States as a path for students to improve their literacy skills. Upon successfully completing these courses, students are deemed proficient in reading and writing and become eligible to participate in an academic program leading to a degree or certificate (Cormier and Bickerstaff, 2019).

Despite the significant role of DevEd, the efficacy of student placement methods, predominantly reliant on standardized entrance assessments such as ACCUPLACER[1], COMPASS[2], and ACT[3], have played a key role in the expansion and reform of DevEd, not only in the United States (King et al., 2017; Kafka, 2018; Zachry Rutschow et al., 2021), but also worldwide (Qian et al., 2020). Studies suggest that these exams misplace, on average, 40% of college-intending students, with a poor correlation between test scores and future college success (Hassel and Giordano, 2015).

At Tulsa Community College (TCC)[4], ACCUPLACER is the primary tool for assessing incoming students' writing proficiency in English (L1). The entrance exam includes the completion of a short essay (300-600 words) on topics like *One's Ability to Change* or *Learning Practical Skills*. Following the submission of students' written productions, ACCUPLACER automatically evaluates and categorizes each

[a] https://orcid.org/0000-0001-8782-8377
[b] https://orcid.org/0000-0003-4603-4364

[1]https://www.accuplacer.org/ (last access: April 5, 2024; all URL in this paper were checked on this date.)
[2]https://www.compassprep.com/practice-tests/
[3]https://www.act.org
[4]https://www.tulsacc.edu/

Da Corte, M. and Baptista, J.
Leveraging NLP and Machine Learning for English (L1) Writing Assessment in Developmental Education.
DOI: 10.5220/0012740500003693

In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 2*, pages 128-140
ISBN: 978-989-758-697-2; ISSN: 2184-5026

essay into a specific tier: DevEd-Level 1, DevEd-Level 2, or College-Level. Every year, over 43% of new students are determined to need a minimum of one DevEd course at this institution based on data reported by its Institutional Research, Reporting, & Analytics department.[5]

For this study, the levels were operationalized as follows: *DevEd-Level 1*: text indicated that development is needed in the overall use of the English language: grammar, spelling, punctuation, and sentence and paragraph structure. *DevEd-Level 2*: text indicated that support is needed in specific areas of the English language, e.g., sentence structure, punctuation, editing, and revising. *College-Level*: text indicated no need for DevEd.

According to The College Board (2022), automatic placement is based on 6 broad linguistic descriptors: (i) *Purpose and Focus*; (ii) *Organization and Structure*; (iii) *Development and Support*; (iv) *Sentence Variety and Style*; (v) *Mechanical Conventions*; and (vi) *Critical Thinking*.

The system's manual definitions of these descriptors (The College Board, 2022) are arguably too abstract (and limited), posing challenges not only to the automatic extraction and assessment of relevant features from texts but also for human annotators to accurately replicate these nuanced intuitions (Da Corte and Baptista, 2024b). Hence, there is a pressing need for a detailed linguistic analysis customized for DevEd, serving as the key motivation for this study.

This study uses NLP tools and Machine-learning (ML) algorithms to assess the effectiveness of various linguistic features sourced from well-known platforms like COH-METRIX and CTAP in a task that classifies texts by proficiency level for student placement in two-level DevEd courses. The research focuses on identifying optimal predictors, feature combinations, and algorithms to enhance placement accuracy (Santos et al., 2021), aiming to improve educational outcomes. By fine-tuning the placement process, more equitable opportunities for linguistically underprepared students can be available, thus reducing educational disparities and supporting fair access to college education (Beaulac and Rosenthal, 2019; Goudas, 2020; Qian et al., 2020).

In view of the limitations and motivations, this study's objectives are twofold: (i) to refine the identification of linguistic features critical for DevEd placement decisions, and (ii) to enhance students' L1 writing proficiency assessment within DevEd.

2 RELATED WORK

Research on enhancing student placement in DevEd courses, particularly for L1 English speakers, has focused on improving classification accuracy (CA) through lexical and syntactic pattern analysis, leveraging Text Mining techniques (Da Corte and Baptista, 2024a).

Pal and Pal (2013) employed the WEKA Machine-learning (ML) platform to classify students into appropriate courses using Naive Bayes, Multilayer Perceptron, and Tree models, achieving a CA of 86.15% and benefiting placement accuracy. Similarly, Filighera et al. (2019) utilized Neural Networks and embeddings to classify texts into 5 reading levels, achieving an accuracy of 81.3% through 5-fold cross-validation.

Using ML, Bujang et al. (2021) developed a multiclass prediction model for course grades, achieving 99.5% accuracy with Random Forest, facilitated by Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). Crossley et al. (2017) analyzed STEM student essays, identifying text variations across disciplines and suggesting subject-specific teaching approaches for DevEd. Subsequent studies like Crossley (2020) emphasized lexical sophistication and syntactic complexity for automatic assessment. Nazzal et al. (2020) advocated for integrating ML with linguistic data in non-standardized assessments to enhance student placement in DevEd.

NLP tools like COH-METRIX[6] (McNamara et al., 2006) and CTAP[7] (Chen and Meurers, 2016) have been pivotal in analyzing linguistic complexity across languages. Leal et al. (2023) adapted cohesion and coherence metrics from Coh-Metrix English to Coh-Metrix (Brazilian) Portuguese, while Okinina et al. (2020) extended CTAP measures to Italian. Akef et al. (2023) used CTAP to assess language proficiency in Portuguese, achieving 76% CA, and highlighted feature selection's role in refining analysis. Recent work by Wilkens et al. (2022) emphasized lexical diversity and dependency counts in French language development assessment.

Identifying more descriptive linguistic features and incorporating them into systems like ACCUPLACER, leveraging NLP and ML algorithms, could enhance skill-level classification, laying the groundwork for this study. This research builds on previous studies that outline a framework for assessing students' linguistic skills, focusing on how outcomes de-

[5]https://www.tulsacc.edu/about-tcc/institutional-research

[6]http://141.225.61.35/CohMetrix2017/

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

termine their course placement and participation in an academic program.

3 METHODS

3.1 Corpus

From a pool of essays written by 290 native speakers enrolled in DevEd courses, 100 essays were randomly selected, ensuring a balanced representation across the two DevEd levels. These essays were produced at the institution's monitored testing center during the 2021-2022 academic year. Essays were written without time constraints or the ability to use editing tools. Despite the modest sample size, it establishes the groundwork for a corpus aimed at documenting the linguistic variety among community college students as they commence their higher education journey.

The samples were extracted from the institution's standardized entrance exam database in plain text format. This process followed the Institution's Review Board (IRB)[8] approved protocols, with the identifier #22-05, focusing on educationally disadvantaged individuals, meeting stringent ethical standards. The main metadata indicated the students' DevEd placement level as assigned by ACCUPLACER. At this point, additional metadata, such as demographics (including gender and race), was not considered.

As presented in Table 1, sample text units were balanced by level but varied in length (number of tokens per text), making the corpus quite unbalanced concerning this metric. A custom Python function from Python's standard libraries was used to tokenize the texts. Punctuation signs were kept as tokens, as punctuation is a potentially good predictor of how proficiently students write at the onset of developing their academic writing skills. No text transformation was used since the upper/lower case distinction may be relevant to model students' behavior in DevEd courses, as they do not utilize capitalization consistently when writing for academic purposes.

Table 1: Original corpus characteristics.

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

To address the length issue, the sampling units were split into segments of 100 words. All sampling units below this threshold were discarded. The result was 94 units from Level 1 and 119 from Level 2. To achieve a balanced corpus across levels, 25 units from Level 2 were excluded through random resampling, resulting in an equal count of 94 units from each level for this analysis.

Results of this trimming process are summarized in Table 2.[9]

Table 2: Corpus characteristics after splitting the text sample units.

Split Sample Text Units	
Level 1	141
Level 2	199
Total	**340**
Text units discarded - Level 1	47
Text units discarded - Level 2	80
Total split samples discarded	**127**
Total balanced split samples for both levels	**188**

3.2 Linguistic Features

A total of 436 linguistic features were extracted from the analyzed sample text units utilizing two distinct analytical tools, COH-METRIX and CTAP, with these features grouped into cluster categories as detailed in Table 3. Specifically, the distribution of these features across the tools is as follows: the COH-METRIX tool accounted for 106 of these features, while CTAP accounted for the remaining 330 features. A detailed description of these features can be found in the documentation of these tools.

These features were supplemented with DevEd-specific (DES) features[10] obtained by the manual annotation of the corpus. In their majority, DES features include features that signal errors and indicate a deviation from proficiency standards; a few reveal patterns that signal proficiency.

The annotation proper of DES features was conducted by two qualified, trained annotators who employed an annotation scheme developed by the authors of this paper (Da Corte and Baptista, 2024a). The 21 most salient features utilized, distributed across 4 textual patterns (feature clusters), are briefly mentioned in Table 4.

To assess the reliability of the annotations, the Krippendorff's Alpha (K-alpha) interrater reliability coefficient was calculated, obtaining a moderate score of k=0.40 (Da Corte and Baptista, 2024a). Given the intricate nature and complexity of the annotation task, this score was considered adequate. Based on this as-

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

[9]The potential bias from the assumed independence of segments in the study was recognized, with the decision to defer addressing it made at this point.

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

Table 3: Feature cluster categories: COH-METRIX & CTAP.

	COH-METRIX	CTAP
Patterns	**Feature Clusters**	**Feature Clusters**
Lexical	Descriptive (e.g., number of tokens) (DESC)	Descriptive (e.g., number of tokens) (DESC)
	Connectives (CONNECT)	Lexical Density (LEXDENS)
		Lexical Richness (LEXRICH)
		Lexical Sophistication (LEXSOPH)
		Lexical Variation (LEXVAR)
Syntactic	Syntactic Complexity (SYNTCOMPLX)	Syntactic Complexity (SYNTCOMPLX)
	Syntactic Pattern Density (SYNTPATTERNDENS)	Number of Syntactic Constituents (NUMSYNTCONST)
	Word Information (WORDINFO)	Number of POS (NUMPOS)
		POS Density (POSDENS)
	Referential Cohesion (REFCOH)	Referential Cohesion (REFCOH)
Discursive	Situation Model (SITMODEL)	-
	Latent Semantic Analysis (LATSEMANALYSIS)	
	Text Easability (TXTEASA)	
Readability	Readability (e.g., Flesch_Kincaid Grade Level) (READ)	-

sessment, a consensual annotation was reached, ultimately retaining 6,495 tags for analysis. All of the tags were systematically accounted for using Python code.

While some of these features partially overlap with those extracted by COH-METRIX and CTAP, others constitute novel contributions to the proficiency assessment field. For example, *Fictional You* (rhetoric, generic representation of a person, using the pronoun *you*) and *Fictional We* (a similar device, but using the pronoun *we*). Multiword expressions (MWE) is another example and was previously investigated in Da Corte and Baptista (2022), which, along with other studies (Laporte, 2018; Kochmar et al., 2020; Pasquer et al., 2020) confirmed that using MWE as lexical features can improve the CA of students in DevEd.

3.3 Experimental Design

Several supervised ML experiments were conducted as part of this study, construed as a classification task, to determine: (i) a selection of the best linguistic features for the task; and (ii) the best-performing ML algorithm. A hardware configuration comprising an 11th Gen Intel(R) Core(TM) i7-1165G7 CPU with a base clock speed of 2.80GHz, complemented by 8.00 GB of RAM, operating on a 64-bit system with an x64-based processor, was used.

The data mining tool ORANGE (Demšar et al., 2013)[11] was selected for analysis and modeling for its usability and the diversity of ML tools and algorithms it makes available. A total of 10 ML algorithms were selected from the set available in ORANGE (in alphabetical order): (i) *Adaptive Boosting* (AdaBoost); (ii) *CN2 Rule Induction* (CN2); (iii) *Decision Tree* (DT) (iv) *Gradient Boosting* (GB); (v) *k-Nearest Neighbors* (kNN); (vi) *Logistic Regression* (LR); (vii) *Naive Bayes* (NB); (viii) *Neural Network* (NN); (ix) *Random Forest* (RF); and (x) *Support Vector Machine* (SVM). The default configuration of these learners was selected. Figure 1 shows the basic workflow adopted for this study.

For the training step and to assess the models, the TEST&SCORE widget was used. Models were assessed using the Classification Accuracy (CA) as the primary evaluation metric, which closely aligns with the task at hand. Precision (Prec) was used as a secondary method to rank the models in the event of *ex aequo* CA values. Given the corpus size, the data was automatically partitioned (DATA SAMPLER) for a 3-fold cross-validation, leaving 2/3 of the corpus for training and 1/3 for testing purposes. The RANK widget was used to assess the discriminative value of each feature for the task. A Confusion Matrix also allowed for a detailed inspection of the results.

Two classification scenarios were devised to assess the impact of text length on the task:

Scenario 1, involves the initial set of full (F) 100 text samples, in their original form, with different text sizes (spanning from 95 to 422 words), and balanced for placement level.

Scenario 2, involves samples split (S) into fragments of 100 words each and then resampled to keep the placement level balanced, as mentioned in Subsection 3.1 and shown in Table 2.

For scenario 2, a new dataset was produced to correspond to the contents of the split essay fragments. The CTAP and COH-METRIX platforms had to be rerun on this new dataset, while the DES features had to be retrieved again.

For each scenario, four experiments were carried out:

Experiment 1, where sample text units were classified using the entire feature sets from COH-METRIX, CTAP, and DES. Due to the availability of compara-

[11] https://orangedatamining.com/

Table 4: DevEd-specific (DES) features summary.

DevED-Specific (DES) Features		
Patterns	**Description**	**Feature Clusters**
Orthographic (ORT)	Patterns representing the foundational language skills needed to represent words and phrases.	Grapheme (addition, omission, transposition, and capitalization) Word split Word boundary merged Punctuation used Contractions
Grammatical (GRAMM)	Patterns evidencing the quality of text production.	Word omitted Word added Word repetition Verb tense Verb disagreement Verb form Pronoun-alternation referential
Lexical & Semantic (LEXSEM)	Patterns contributing to the structuring of a writer's discourse.	Slang Multiword expressions (MWE) Word precision Mischosen preposition Connectives
Discursive (DISC)	Patterns exhibiting the writer's ability to produce extended discourse.	Fictional '*we*' Fictional '*you*' Argumentation with reason Argumentation with example

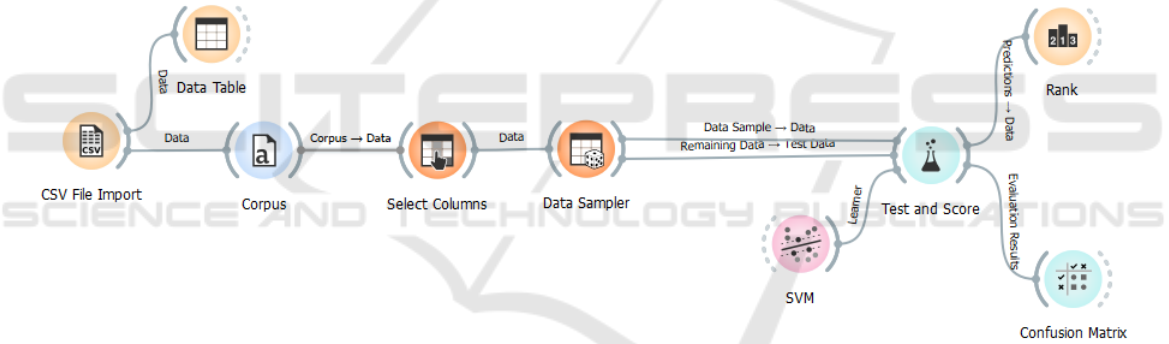


Figure 1: ORANGE workflow setup. The SVM algorithm is displayed merely as a representative of the chosen learners.

ble data, this experiment serves as the *baseline* for the 3 next experiments.

Experiment 2, classified the text samples based on the top 11 more discriminative features, as indicated by the RANK widget for each feature set; two ranking measures were compared: the *Information Gain* and *Chi-square* (χ^2) scoring methods.

Experiment 3, using a one-out approach, the classification involved removing one feature cluster at a time. These clusters have been presented in Tables 3 and 4. The goal here was to measure the magnitude of the decrease in the CA of the ML algorithms. For the analysis of the results, the following guiding principle was adopted: the larger the decrease, the greater the significance of the feature cluster.

Experiment 4, consisted in classifying the text sample units by aggregating features from the three distinct sets (COH-METRIX, CTAP, and DES) and subsequently identifying the most discriminative ones using the *Information Gain* ranking method, which is a common method used for feature selection.

4 RESULTS

The results from Experiments 1 through 4 are detailed in this section, providing a comparison of classification accuracies across different experimental setups and feature analysis methods. The dataset with the respective scores (ratios) for all 457 linguistic features mentioned in Section 3.2 can be found on Da Corte and Baptista (2024c).[12]

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

Experiment 1

Table 5 presents the CA scores of the 10 different ML models introduced in Section 3.3, applied to both full (F) and split (S) scenarios and utilizing using different feature sets (COH-METRIX, CTAP, and DES). Notably, for COH-METRIX, GB and LR achieved the highest CA scores for F (0.697) and S (0.616) scenarios, respectively. With the CTAP feature set, RF outperformed COH-METRIX for the F scenario, yielding a CA of 0.727, whereas NN was the best-performing learning algorithm in the S scenario with a CA of 0.624. Using the DES feature set, CN2 and NB performed comparatively to the classification scores obtained with COH-METRIX, achieving accuracies of 0.652 (CN2) for the full scenario and 0.640 (NB) for split samples. Regarding the S scenario, the NB with the DES feature set achieved the highest CA score.

As previously mentioned, this experiment establishes the baseline for this study, aiming to enhance the CA beyond the 0.727 benchmark set by RF (in the F scenario). This benchmark is relatively high and correlates to the fact that RF is often recognized for its efficacy in ML applications, particularly in the context of writing analysis (Huang, 2023).

Experiment 2

Two feature ranking methods, Information Gain and χ^2, were used to identify the top 11 best-performing features. The two ranking methods produced very different results. To quantify this discrepancy between the ranking methods, the Spearman Rank Correlation coefficient was calculated, which resulted in a moderate correlation (Schober et al., 2018) score of ρ = 0.575. In general, the top 11 features selected using Information Gain yielded better CA results for most models and in both scenarios than those produced by χ^2, and thus, chosen for feature selection. Due to space limitations, these results are not presented here. The outcome of this selection process is detailed in Table 6. For each feature source, specifically COH-METRIX and CTAP, the descriptions provided by the respective feature extraction platforms were utilized.

Table 7 presents the differences in CA values between Experiment 2 and Experiment 1 (baseline) for F and S scenarios. Positive values indicate an increase in CA (from the baseline), while negative values (-) indicate a decrease. The largest increase per ML algorithm's CA based on COH-METRIX, CTAP, and DES feature sets is in bold, while the largest decrease is italicized.

A notable increase in CA of nearly 14% is observed for the NB and GB models on the full scenario. This improvement was achieved by employing only the top 11 features identified through Information Gain from the COH-METRIX and CTAP feature sets. In contrast, with the DES feature set, the increase in accuracy in the full scenario was comparatively smaller, at 6.1% for DT, which is less than half of the improvement observed with the previous models. In the S scenario, NB demonstrated a 16% accuracy increase with CTAP features, likely due to the uniform size of text sample units. This model also showed nearly a 10% improvement with COH-METRIX features. Meanwhile, GB exhibited a more modest increase of 4.8% with the DES feature set.

The largest performance decline was observed with the kNN model for both F and S scenarios, showing decreases of 7.6% and 7.2%, respectively, when employing the top 11 COH-METRIX features. When the top CTAP features were used, the LR model's accuracy slightly decreased by less than 5% for the full scenario. However, in the S scenario, the performance across all models increased. For the DES features, GB and NN experienced a decline of 6.1% (F) and 6.4% (S), respectively.

Results, as presented in Table 7, indicate that feature selection generally enhances the performance of the models, with the exception of DES when applied to the full scenario. However, based on the information included in Table 6, what can be inferred from the selected features from each feature set is very limited, as they correspond to very disparate properties, e.g., *Sentence length, number of words, mean*; *Flesch Reading Ease*; *Number of tokens*; *Number of POS feature: existential there tokens*; *Mischosen preposition*; *MWE*. To gain a better insight into the predictive impact of these feature sets, they were clustered by types, which is the purpose of Experiment 3.

Experiment 3

In this experiment, features were clustered by type (within their respective platform), and each cluster was sequentially removed. The models tested, along with their ORANGE configuration, remained as introduced initially and presented in Figure 1. To interpret the results, positive values in the classification experiment denote an improved CA when the cluster is removed, while negative values indicate a hindered classification. The aim is to pinpoint the most crucial feature clusters for the task, particularly focusing on those whose removal significantly impacts classification. The largest decrease in CA per ML algorithm is highlighted in bold, while the largest decrease per cluster is italicized.

COH-METRIX

First, Table 8 presents the changes in CA values for the full (F) and split (S) scenarios, as com-

Table 5: Experiment 1: Classification Accuracy (CA): full (F) vs. split (S) scenarios using COH-METRIX, CTAP, and DES features.

Features	COH-METRIX		CTAP		DES	
Model	CA (F)	CA (S)	CA (F)	CA (S)	CA (F)	CA (S)
AdaBoost	0.621	0.560	0.606	0.560	0.455	0.576
CN2	0.667	0.584	0.591	0.560	**0.652**	0.520
DT	0.606	0.568	0.606	0.504	0.515	0.616
GB	**0.697**	0.584	0.545	0.528	0.561	0.592
kNN	0.682	0.592	0.621	0.504	0.576	0.560
LR	0.652	**0.616**	0.652	0.568	0.561	0.464
NB	0.652	0.576	0.712	0.552	0.591	**0.640**
NN	0.636	0.560	0.606	**0.624**	0.515	0.632
RF	0.652	0.560	**0.727**	0.512	0.530	0.632
SVM	0.576	0.576	0.712	0.544	0.561	0.552

Table 6: Experiment 2: Top 11-ranked features, per feature source, ranked by Information Gain method.

COH-METRIX	CTAP	DES
Sentence length, number of words, $\bar{X}$	Lexical sophistication: easy word types (NGSL)	Argumentation with example
Flesch-Kincaid grade level	Syntactic complexity feature: prepositional phrases per sentence	Word omitted
Left embeddedness, words before main verb, $\bar{X}$	Number of word types with more than 2 syllables	Mischosen preposition
Word count, number of words	Number of tokens	Word precision
Flesch Reading Ease	Number of tokens with more than 2 syllables	Grapheme
Negative connectives incidence	Number of POS feature: adverb lemma types	Word repetition
Paragraph length, number of sentences in a paragraph, σ	Number of POS feature: existential there tokens	Verb disagreement
Sentence syntax similarity, all combinations, across paragraphs, $\bar{X}$	Lexical sophistication: easy lexical types (NGSL)	Multiword Expressions
Lexical diversity, type-token ratio, content word lemmas	Lexical sophistication: easy lexical tokens (NGSL)	Argumentation with reason
Text easability PC syntactic simplicity, z score	Number of POS feature: preposition types	Pronoun-alternation referential
Text Easability PC Syntactic simplicity, percentile	Number of syntactic constituents: postnominal noun modifier	Punctuation used

Table 7: Experiment 2: Classification Accuracy (CA) differences from baseline (Experiment 1) for full (F) vs. split (S) scenarios using top 11 features ranked by Information Gain. Baseline: 0.727 (F) and 0.640 (S).

Features	COH-METRIX		CTAP		DES	
Model	CA (F)	CA (S)	CA (F)	CA (S)	CA (F)	CA (S)
AdaBoost	-0.015	0.008	0.000	0.064	0.045	0.040
CN2	0.000	-0.024	0.121	0.112	-0.031	0.016
DT	-0.015	0.032	0.015	0.120	**0.061**	0.008
GB	-0.045	0.040	**0.137**	0.112	*-0.061*	**0.048**
kNN	*-0.076*	*-0.072*	0.091	0.104	-0.031	0.032
LR	0.015	-0.024	*-0.046*	0.072	0.000	0.000
NB	**0.136**	**0.096**	0.030	**0.160**	0.045	0.008
NN	0.016	0.080	0.121	0.024	-0.015	*-0.064*
RF	0.045	0.024	0.015	0.128	-0.060	-0.032
SVM	0.060	0.032	0.046	0.096	0.030	0.016

pared to the baseline, using COH-METRIX features. Within the F scenario analysis, significant decreases in CA scores were observed when holding out the Descriptive (DESC), Syntactic Complexity (SYNTCOMPLX), and Word Information (WORDINFO) clusters. DESC focuses on formal text properties like sentence length and word count, while SYNTCOMPLX focuses on syntactic aspects such as left embeddedness and sentence syntax similarity. WORDINFO includes cognitive features associated with language development, like age of acquisition and familiarity for content words.

The impact of holding out these clusters varied across learning models, with notable decreases observed with the kNN, LR, and CN2 algorithms. For example, the kNN model experienced a large decrease of nearly 23% when the DESC cluster was removed. At the same time, LR saw a decrease of almost 11% with the removal of WORDINFO, and CN2 experienced a decrease of 9.1% with the removal of SYNTCOMPLEX. All of these clusters belong to lexical and syntactic patterns.

In contrast, the S scenario showed improvements in performance for many ML models when certain feature clusters were removed, suggesting that these clusters may hinder the classification task when included. The clusters leading to the most considerable increases in CA included Syntactic Pattern Density (SYNTPATTERNDENS), Situational Model (SITMODEL), Connectives (CONNECT), and Referential Cohesion (REFCOH), each contributing to different linguistic aspects. These clusters are associated with syntactic and discursive patterns.

Results varied depending on the learner used. For instance, the NN model showed an average improvement of 13.7% when the four mentioned clusters were removed, with the highest improvement of 18.2% attributed to the removal of SYNTPATTERNDENS.

Table 8: Experiment 3: Changes in Classification Accuracy (CA) for full (F) vs. split (S) scenarios using COH-METRIX features with one-out feature cluster removal. Baseline: 0.727 (F) and 0.640 (S).

		COH-METRIX									
		Models									
Holdout Clusters	**CA**	**AdaBoost**	**CN2**	**DT**	**GB**	**kNN**	**LR**	**NB**	**NN**	**RF**	**SVM**
DESC	F	0.015	-0.031	0.000	**-0.076**	***-0.227***	-0.061	**-0.061**	0.016	**-0.061**	0.015
	S	0.092	0.052	0.038	0.068	*-0.137*	-0.025	**-0.061**	0.076	0.107	0.015
CONNECT	F	**-0.015**	0.000	**-0.015**	-0.015	*-0.061*	0.000	0.015	0.031	0.000	0.030
	S	0.046	0.083	0.023	0.083	0.029	0.036	0.015	0.137	0.107	0.030
SYNTCOMPLX	F	0.061	***-0.091***	0.046	-0.015	0.000	0.000	-0.031	0.016	-0.031	0.000
	S	0.107	-0.008	0.084	0.098	0.090	0.036	-0.031	0.061	0.031	0.000
SYNTPATTERNDENS	F	0.015	*-0.031*	**-0.015**	0.015	0.015	0.045	-0.016	0.031	0.030	0.000
	S	0.076	0.052	0.023	0.128	0.105	0.096	*-0.016*	0.182	0.001	0.000
WORDINFO	F	0.015	-0.061	0.015	**-0.076**	-0.015	***-0.107***	0.045	0.046	0.030	0.060
	S	0.046	0.037	0.053	0.037	0.075	*-0.040*	0.045	0.122	0.061	0.060
REFCOH	F	0.046	*-0.031*	0.000	0.000	0.000	0.000	0.000	0.076	-0.016	0.045
	S	0.107	0.052	0.038	0.113	0.090	0.036	0.000	0.152	0.046	0.045
SITMODEL	F	0.046	0.000	0.015	0.000	0.000	*-0.016*	0.015	0.031	0.000	0.015
	S	0.061	0.083	0.053	0.113	0.090	0.020	0.015	0.076	0.182	0.015
LATSEMANALYSIS	F	0.031	-0.031	0.030	*-0.045*	0.000	0.000	0.045	0.031	0.030	0.030
	S	0.107	0.052	0.068	0.068	0.090	0.036	0.045	0.107	0.001	0.030
TXTEASA	F	0.046	-0.031	0.000	0.000	*-0.061*	0.075	0.015	0.000	-0.046	0.030
	S	0.107	0.052	0.038	0.113	0.029	0.111	0.015	0.076	-0.015	0.030
READ	F	0.076	*-0.061*	0.030	-0.030	-0.015	-0.031	0.030	**-0.030**	0.060	0.000
	S	0.137	*-0.069*	0.053	0.083	0.075	0.005	0.030	0.107	-0.015	0.000

Similarly, GB increased by almost 11% on average, with the highest improvement of 12.8% attributed to the removal of SYNTPATTERNDENS. RF also showed notable improvements, with a high increase of 18.2% when the SITMODEL cluster was removed.

CTAP

Next, Table 9 presents the CA values for the same two scenarios with CTAP features. Within the F scenario analysis, significant decreases in CA scores were observed when the holdout strategy included the Lexical Richness (LEXRICH), Lexical Variation (LEXVAR), Number of Part-of-speech (NUMPOS), and Referential Cohesion (REFCOH) clusters. LEXRICH and LEXVAR focus on lexical patterns, NUMPOS includes adverb lemma types and existential *there* tokens, and REFCOH encompasses local lexical overlap and noun overlap, all of which fall under syntactic patterns.

The most notable findings were with the RF model, which showed a decrease in performance when all four clusters were removed. Specifically, removing LEXRICH led to the largest drop of nearly 11%, while removing LEXVAR, NUMPOS, and REFCOH caused a 9.1% decrease each. However, the CN2 model exhibited a remarkable 15.1% increase in accuracy when LEXRICH was excluded, while other clusters did not affect classification. The GB model saw a 6.1% accuracy improvement when LEXVAR was removed, 1.6% for LEXRICH, 3.1% NUMPOS, and no impact for REFCOH. This asymmetry in the performance of the models requires careful interpretation of the results and cannot be directly translated into a choice of the best-performing feature clusters in this task. This will be the object of subsequent studies. Also, it is pertinent to note that this asymmetry has not been observed in such an expressive way with the COH-METRIX feature clusters.

For the S scenario, in addition to LEXRICH, LEXVAR, and NUMPOS, Lexical Density (LEXDENS) and Lexical Sophistication (LEXSOPH) clusters were considered for their impact on the ML model's performance. LEXDENS includes features like modals per word frequency, while LEXSOPH comprises simple word presence and lexical types from the New General Service List (NGSL). The AdaBoost model saw a significant decrease of nearly 13% in accuracy when NUMPOS was excluded, and CN2 exhibited a uniform decrease of nearly 9% across several clusters. However, for the RF model, removing LEXSOPH and NUMPOS clusters resulted in a comparatively modest average decrease in accuracy of nearly 7%

DES

The one-out cluster removal strategy was applied to the DES features as a last step in this experiment. Table 10 presents the changes in CA values for the same two scenarios, compared to the baseline. Within the F scenario analysis, significant decreases in CA scores occurred when the holdout strategy included Orthographic (ORT) and Grammatical (GRAMM) patterns, with ORT impacting multiple models. The features within the DES clusters have been previously introduced in Table 4.

The most notable findings involved a decrease in CA ranging from 15.2% for GB, 12.1% for kNN, to almost 11% for NN when ORT was removed. Con-

Table 9: Experiment 3: Changes in Classification Accuracy (CA) for full (F) vs. split (S) scenarios using CTAP features with one-out cluster removal. Baseline: 0.727 (F) and 0.640 (S).

		CTAP									
		Models									
Holdout Clusters	**CA**	**AdaBoost**	**CN2**	**DT**	**GB**	**kNN**	**LR**	**NB**	**NN**	**RF**	**SVM**
DESC	F	0.030	0.000	0.000	0.076	0.000	*-0.046*	-0.015	-0.015	-0.015	0.000
	S	*-0.040*	0.008	-0.016	0.000	-0.008	-0.024	0.016	0.000	0.040	-0.008
LEXDENS	F	0.000	0.121	0.046	0.016	0.000	0.000	0.000	0.015	*-0.075*	0.000
	S	-0.080	***-0.088***	0.024	0.008	0.000	0.000	0.000	**-0.056**	0.056	0.000
LEXRICH	F	0.015	0.151	3.000	0.016	0.000	0.000	0.000	0.015	***-0.106***	**-0.030**
	S	-0.032	***-0.088***	0.048	0.040	0.000	-0.008	0.024	0.024	-0.016	0.048
LEXSOPH	F	0.000	0.106	0.000	0.016	0.046	***-0.076***	0.000	0.061	-0.015	-0.030
	S	*-0.088*	***-0.088***	0.016	0.024	**-0.032**	-0.024	0.000	0.016	0.064	0.008
LEXVAR	F	-0.030	0.000	0.000	0.061	0.000	0.000	0.000	0.046	*-0.091*	0.000
	S	-0.024	***-0.088***	0.016	0.032	0.000	0.000	0.008	0.008	0.040	-0.016
SYNTCOMPLX	F	-0.015	0.000	***-0.061***	0.000	0.000	0.060	0.015	0.046	-0.060	0.000
	S	-0.040	0.040	0.008	-0.008	0.008	*-0.040*	**-0.024**	-0.024	0.016	0.040
NUMSYNTCONST	F	0.000	0.000	-0.015	-0.015	0.000	0.015	0.000	0.046	*-0.030*	0.000
	S	-0.024	*-0.072*	0.032	**-0.032**	0.000	-0.008	0.024	-0.032	0.032	-0.024
NUMPOS	F	0.046	0.000	-0.015	0.031	0.000	-0.031	-0.015	0.030	*-0.091*	-0.015
	S	***-0.128***	**-0.088**	-0.024	0.000	0.000	0.000	0.024	0.008	0.072	0.016
POSDENS	F	0.030	0.000	0.000	0.076	0.000	0.000	0.000	0.000	0.000	0.000
	S	-0.032	*-0.072*	0.008	0.008	0.000	0.000	0.016	-0.040	0.056	0.040
REFCOH	F	-0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.046	*-0.091*	0.000
	S	-0.040	*-0.072*	0.032	0.032	0.000	0.000	0.000	0.008	0.056	-0.016

Table 10: Experiment 3: Changes in Classification Accuracy (CA) for full (F) vs. split (S) scenarios using DES features with one-out cluster removal. Baseline: 0.727 (F) and 0.640 (S).

		DES									
		Model									
Holdout Clusters	**CA**	**AdaBoost**	**CN2**	**DT**	**GB**	**kNN**	**LR**	**NB**	**NN**	**RF**	**SVM**
ORT	F	-0.031	-0.046	-0.045	***-0.152***	**-0.121**	0.000	**-0.076**	**-0.106**	-0.060	0.000
	S	-0.032	-0.008	-0.048	-0.024	-0.032	**-0.008**	-0.016	*-0.096*	-0.064	-0.008
GRAMM	F	0.121	***-0.122***	0.015	0.151	0.000	0.000	0.091	0.030	0.152	0.030
	S	-0.016	*-0.064*	-0.008	-0.008	-0.056	0.000	-0.048	-0.040	*-0.064*	0.000
LEXSEM	F	0.075	*-0.076*	-0.015	0.060	-0.031	0.000	-0.046	0.000	0.015	0.015
	S	***-0.144***	-0.080	**-0.136**	-0.112	-0.008	0.000	-0.016	-0.104	**-0.112**	**-0.088**
DISC	F	0.075	0.015	-0.015	-0.031	*-0.076*	0.000	0.000	-0.030	0.106	-0.046
	S	0.040	-0.016	-0.024	0.000	*-0.096*	0.000	-0.016	-0.064	-0.064	-0.048

versely, removing GRAMM only affected the CN2 model, decreasing its performance by 12.2%, while AdaBoost, GB, NB, NN, and RF improved their accuracy by almost 11% on average. Both ORT and GRAMM features are indicative of formal correction, making them particularly relevant to the classification task, especially in the F scenario.

For the S scenario, LEXSEM had the highest impact on CA scores. Five models, including AdaBoost, DT, GB, RF, and NN, experienced a performance deterioration ranging from 10.4% to 14.4% when LEXSEM was removed. As LEXSEM relates to the lexicon used, it has a significant impact on this task's scenario and is less affected by text length.

Experiment 4

The final experiment combined features from three sources: COH-METRIX, CTAP, and DES. It then used the Information Gain ranking method to pinpoint the most discriminative features for both full (F) and split (S) scenarios, before evaluating the performance of ML models. As models were tested, features were added in packs of 10 at a time, prioritized by their Information Gain scores, until reaching asymptotic results. The same suite of ML algorithms employed in prior experiments was used. The higher-ranking selected features (30) and their corresponding Information Gain scores are delineated in Table 11.

Within the highest-ranked features, most come from CTAP (70%), followed by COH-METRIX (23%). On a smaller scale (7%), the presence of two DES features was noted (Verb Disagreement, Information Gain: 0.075; Multiword Expressions (MWE): Information Gain: 0.060) in the 7th and 14th place. The Verb Disagreement is a syntactic feature that is arguably difficult to obtain automatically, while the MWE has seldom been mentioned in the literature concerning readability and/or proficiency estimation studies. Information Gain scores within these 30 features ranged from 0.120 to 0.049. Two other DES

Table 11: Experiment 4: Combined top-ranked 30 features from COH-METRIX, CTAP, and DES, ranked by Information Gain scores.

Rank	Feature	Description	Info. gain
1	CTAP	POS Density Feature: Particle	0.120
2	CTAP	Lexical Richness: Sophisticated Noun Type Ratio (NGSL)	0.114
3	CTAP	Syntactic Complexity Feature: Prepositional Phrases per Sentence	0.105
4	CTAP	Lexical Richness: Sophisticated Noun Ratio (NGSL)	0.087
5	CTAP	Syntactic Complexity Feature: Complex Prepositional Phrases per Sentence	0.082
6	COH-METRIX	CELEX word frequency for content words, $\bar{X}$	0.075
7	DES	Verb Disagreement	0.075
8	CTAP	Lexical Sophistication: Sophisticated noun tokens (NGSL)	0.071
9	CTAP	Number of POS Feature: Singular or mass noun Types	0.066
10	CTAP	Number of POS Feature: Particle Tokens	0.066
11	COH-METRIX	Hypernymy for nouns, $\bar{X}$	0.064
12	CTAP	Mean Sentence Length in Letters	0.062
13	CTAP	Mean Sentence Length in Syllables	0.060
14	DES	Multiword Expressions	0.060
15	CTAP	Syntactic Complexity Feature: Mean Length of Complex T-unit	0.057
16	CTAP	POS Density Feature: Existential There	0.057
17	CTAP	Number of POS Feature: Possessive ending Tokens	0.054
18	COH-METRIX	Text Easability PC Narrativity, percentile	0.054
19	CTAP	POS Density Feature: Possessive Ending	0.053
20	COH-METRIX	Sentence length, number of words, σ	0.053
21	CTAP	POS Density Feature: Modal Verb	0.052
22	COH-METRIX	Stem overlap, all sentences, binary, $\bar{X}$	0.051
23	CTAP	Syntactic Complexity Feature: Complex T-unit per Sentence	0.051
24	CTAP	Lexical Richness: Easy Lexical Type Ratio (NGSL)	0.051
25	COH-METRIX	Sentence length, number of words, $\bar{X}$	0.051
26	COH-METRIX	Ratio of intentional particles to intentional verbs	0.051
27	CTAP	Number of POS Feature: Existential there Tokens	0.050
28	CTAP	Syntactic Complexity Feature: Sentence Complexity Ratio	0.050
29	CTAP	Lexical Sophistication: Easy noun types (NGSL)	0.050
30	CTAP	Number of Syntactic Constituents: Verb Phrase	0.049

features ranked 42nd (Slang) and 46th (Mischosen Preposition), both from the lexical and semantic pattern clusters.

Table 12 illustrates the impact of this feature selection on the predictive accuracy of the employed ML algorithms. CA scores are highlighted in bold to denote the highest scores achieved with varying numbers of features (Ft) - 10Ft to 100Ft. Additionally, scores that exceed the benchmark CA of 0.727, established in Experiment 1, are italicized for each model and feature set. Figure 2 depicts the outcomes of Experiment 4 for a more in-depth evaluation of the results obtained here.

Several algorithms, namely the CN2, DT, GB, LR, and SVM, underperformed relative to the baseline. Notably, the CN2 algorithm consistently registered a CA of 0.561, showing no improvement with the increase in feature count, thus indicating it is not suitable for the complex DevEd classification task devised for this study.

Conversely, AdaBoost looked like a promising model by exceeding the baseline CA with 70 Ft; however, results in CA tend to decrease upon adding further features. A similar trend was observed with the kNN, performing barely over the baseline with 80 Ft but quickly decreasing its performance with the addition of more features. NN achieved a notable CA of 0.758 with both 60 and 70 Ft, yet it showed no further improvements beyond this point.

The RF showed to be a fast learning model and a promising one for this type of classification task. The model performed consistently when both 30 and 40 features were added, exhibiting a CA score of 0.742 in both instances. Beyond this point, the model's performance deteriorated considerably.

Among the algorithms, NB stood out as the fastest learning model and the one that consistently performed the best throughout the experiment. With 30 Ft, the model achieved a CA of 0.773 —an almost 5% enhancement over the baseline. Its performance reached an asymptotic line at a CA of 0.788 with 50 Ft, marking a notable 9.1% improvement from the baseline. The model continued performing above the baseline as more features were added, reaching a peak CA score of 0.818 at 60 Ft. As more features were added, the model performed consistently within a CA range of 0.773 and 0.788.

When the experiment was conducted within the S scenario, only one model, the NN, performed above the baseline (0.640) with a CA of 0.656 (with 10Ft), which is only a 1.6% improvement. As more features were added, scores deteriorated substantially, with accuracy scores ranging between 0.432 and 0.480. Therefore, these scores were discarded.

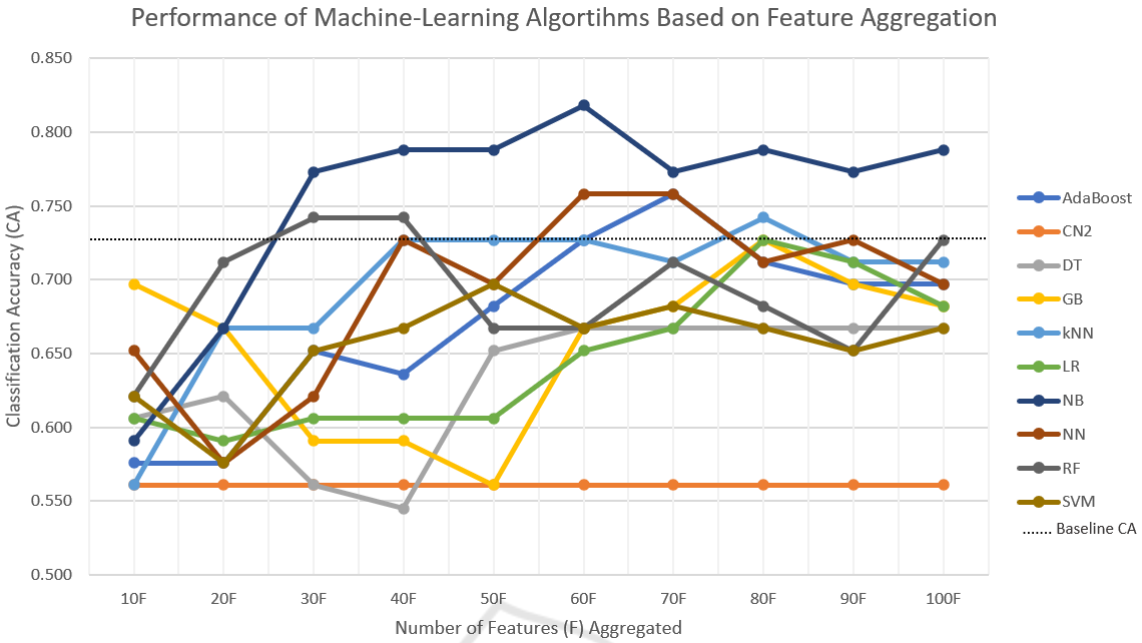


Figure 2: Experiment 4: Machine-learning algorithms performance.

Table 12: Experiment 4: Classification Accuracy (CA) for full (F) scenario using a combination of feature sets (Ft), in packs of 10, based on Information Gain.

	Classification Accuracy (CA) Scores									
Model	**10Ft**	**20Ft**	**30Ft**	**40Ft**	**50Ft**	**60Ft**	**70Ft**	**80Ft**	**90Ft**	**100Ft**
AdaBoost	0.576	0.576	0.652	0.636	0.682	0.727	*0.758*	0.712	0.697	0.697
CN2	0.561	0.561	0.561	0.561	0.561	0.561	0.561	0.561	0.561	0.561
DT	0.606	0.621	0.561	0.545	0.652	0.667	0.667	0.667	0.667	0.667
GB	**0.697**	0.667	0.591	0.591	0.561	0.667	0.682	0.727	0.697	0.682
kNN	0.561	0.667	0.667	0.727	0.727	0.727	0.712	0.742	0.712	0.712
LR	0.606	0.591	0.606	0.606	0.606	0.652	0.667	0.727	0.712	0.682
NB	0.591	0.667	***0.773***	***0.788***	***0.788***	***0.818***	***0.773***	***0.788***	***0.773***	***0.788***
NN	0.652	0.576	0.621	0.727	0.697	0.758	0.758	0.712	0.727	0.697
RF	0.621	**0.712**	0.742	0.742	0.667	0.667	0.712	0.682	0.652	0.727
SVM	0.621	0.576	0.652	0.667	0.697	0.667	0.682	0.667	0.652	0.667

5 CONCLUSIONS AND FUTURE WORK

This study aimed to address two primary objectives: (i) the refinement of linguistic feature identification crucial to DevEd placement decisions, and (ii) the improvement of first language (L1) writing proficiency assessment within DevEd contexts.

A total of 436 linguistic features were extracted from COH-METRIX and CTAP and supplemented with 21 DES features systematically vetted and tested through a rigorous quality assurance process. A total of 4 supervised ML experiments were conducted within two scenarios (F and S essays) to determine the best linguistic features for the task and the best-performing ML algorithm using ORANGE Text Mining platform. Due to the availability of comparable data, a baseline (0.727, F samples scenario) was set. In general, full samples tend to produce higher accuracy results than when the samples are split.

Improvements in the models' performance were noted. A notable increase in CA of nearly 14% is observed for the NB and GB models on the full scenario in Experiment 2, employing only the top 11 features identified through Information Gain from the COH-METRIX and CTAP feature sets. When the holdout cluster strategy was applied in Experiment 3, accuracy performance varied across learning models, with notable decreases (meaning that removing the clusters hindered the classification task) observed with the kNN, LR, and CN2 algorithms. Clusters were distributed among lexical, syntactic, and discursive pat-

terns, which seems to correlate with some of the patterns reported by the current literature.

The NB evidenced an impressive increase of 9.1% (from the baseline), noted in Experiment 4, when a combination of 60 features from COH-METRIX, CTAP, and DES were used. This ML algorithm, known for its simplicity and adaptability to classification tasks, appeared as a fast learner with a combination of 30 features (21 from CTAP, 7 from COH-METRIX, and 2 DES), yielding a CA of 0.773, and the one that consistently performed. The best performance of this model, with 0.818 in its CA, however, was attained when 60 features were added. While the best-performing features were from CTAP and COH-METRIX, novel features devised explicitly for DevEd purposes, DES, ranked within the top 15.

The limited size of the corpus utilized in this study is recognized. The next phase of this study includes the expansion of the corpus to a more sufficiently robust size, using text samples collected during the 2023-2024 academic year. This expansion will involve the integration of features identified as crucial for enhancing the accuracy of our ML-based classification algorithms. The more accurate the ML-based estimations of classification, the more accurate the placement of students in a DevEd level that closely matches their current writing proficiency levels.

Additionally, large foundational models, specifically those built on Generative Pre-trained Transformer (GPT) technology, will be explored to generate sample texts that align with college-level writing standards and thus test the generalization power on this artificial data of the features and models discussed in this study.

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 profound gratitude to the dedicated annotators who participated in this task and the IT team whose expertise made the systematic analysis of the linguistic features presented in this paper possible. Their meticulous work and innovative approach have been instrumental in advancing our research.

REFERENCES

Akef, S., Mendes, A., Meurers, D., and Rebuschat, P. (2023). Linguistic complexity features for automatic Portuguese readability assessment. In *XXXIX Encontro Nacional da Associação Portuguesa de Linguística, Covilhã, Portugal, October 26–28, 2023, Proceedings 14*, pages 103–109. Associação Portuguesa de Linguística.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, 60:1048–1064.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., and Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, 9:95608–95621.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, X. and Meurers, D. (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.

Cormier, M. and Bickerstaff, S. (2019). Research on Developmental Education Instruction for Adult Literacy Learners. *The Wiley Handbook of Adult Literacy*, pages 541–561.

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

Crossley, S. A., Russell, D. R., Kyle, K., and Römer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, pages 48–81.

Da Corte, M. and Baptista, J. (2022). A phraseology approach in developmental education placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86.

Da Corte, M. and Baptista, J. (2024a). Charting the linguistic landscape of developing writers: an annotation scheme for enhancing native language proficiency. In *Proceedings of the 2024 joint International Conference on Computational Linguistics, Language Resources and Evaluation – LREC-COLING, 20-25 May, 2024, Turin, Italy*, page to appear.

Da Corte, M. and Baptista, J. (2024b). Enhancing writing proficiency classification in developmental education: the quest for accuracy. In *Proceedings of the 2024 joint International Conference on Computational Linguistics, Language Resources and Evaluation - LREC-COLING, 20-25 May, 2024, Turin, Italy*, page to appear.

Da Corte, M. and Baptista, J. (2024c). Linguistic features analysis in a developmental education context:

A comparative study of full and split sample text units. GitLab repository.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 335–348. Springer.

Goudas, A. M. (2020). Measure twice, place once: Understanding and applying data on multiple measures for college placement. http://communitycollegedata.com /wp-content/uploads/2020/03/2020MultipleMeasure sNOSSPreconfWksp.pdf.

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

Huang, Z. (2023). An intelligent scoring system for english writing based on artificial intelligence and machine learning. *International Journal of System Assurance Engineering and Management*, pages 1–8.

Kafka, T. (2018). Student assessment. In Flippo, R. F. and Bean, T. W., editors, *Handbook of College Reading and Study Strategy Research*, pages 326–339. Routledge, 3 edition.

King, J. B., McIntosh, A., Bell-Ellwanger, J., Schak, O., Metzger, I., Bass, J., McCann, C., and English, J. (2017). Developmental Education: Challenges and Strategies for Reform. *US Department of Education, Office of Planning, Evaluation and Policy Development*.

Kochmar, E., Gooding, S., and Shardlow, M. (2020). Detecting multiword expression type helps lexical complexity assessment. *arXiv preprint arXiv:2005.05692*.

Laporte, E. (2018). Choosing features for classifying multiword expressions. In Sailer, M. and Markantonatou, S., editors, *Multiword expressions: In-sights from a multi-lingual perspective*, pages 143–186. Language Science Press, Berlin.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, pages 1–38.

McNamara, D. S., Ozuru, Y., Graesser, A. C., and Louwerse, M. (2006). Validating CoH-Metrix. In *Proceedings of the 28th annual Conference of the Cognitive Science Society*, pages 573–578.

Nazzal, J. S., Olson, C. B., and Chung, H. Q. (2020). Differences in Academic Writing across Four Levels of Community College Composition Courses. *Teaching English in the Two Year College*, 47(3):263–296.

Okinina, N., Frey, J.-C., and Weiss, Z. (2020). Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020*, pages 7123–7131.

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

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

Qian, L., Zhao, Y., and Cheng, Y. (2020). Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research*, 58(4):771–790.

Santos, R., Rodrigues, J., Branco, A., and Vaz, R. (2021). Neural text categorization with transformers for learning portuguese as a second language. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 715–726. Springer.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

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

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K. P., and François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Zachry Rutschow, E., Edgecombe, N., and Bickerstaff, S. (2021). A Brief History of Developmental Education Reform.

Refining English Writing Proficiency Assessment and Placement in Developmental Education Using NLP Tools and Machine Learning

Miguel Da Corte[1,2][a] and Jorge Baptista[1,2][b]
[1]*University of Algarve, Faro, Portugal*
[2]*INESC-ID Lisboa, Lisbon, Portugal*

Keywords: Developmental Education (DevEd), Automatic Writing Assessment Systems, Natural Language Processing (NLP), Machine-Learning Models.

Abstract: This study investigates the enhancement of English writing proficiency assessment and placement for Developmental Education (DevEd) within U.S. colleges using Natural Language Processing (NLP) and Machine Learning (ML). Existing automated placement tools, such as ACCUPLACER, often lack transparency and struggle to identify nuanced linguistic features necessary for accurate skill-level classification. By integrating human-annotated linguistic features, this study aims to contribute to equitable and transparent placement systems that better address students' academic needs, reducing misplacements and their associated costs. For this study, a 300-essay corpus was compiled and manually annotated with a refined set of 11 DevEd-specific (DES) features, alongside 328 linguistic features automatically extracted from CTAP and 106 via COH-METRIX. Supervised ML algorithms were used to compare ACCUPLACER-generated classifications with human ratings, assessing classification accuracy and identifying predictive features. This analysis revealed gaps in ACCUPLACER's classification capabilities. Experimental results showed that models incorporating DES features improved classification accuracy, with Naïve Bayes (NB) and Support Vector Machine (SVM) achieving scores up to 80%. The refined features presented and methodology offer actionable insights for faculty and institutions, potentially contributing to more effective DevEd course placements and targeted instructional interventions.

1 INTRODUCTION AND OBJECTIVES

Developmental Education (DevEd) courses play a crucial role in equipping students who are not yet prepared for college-level work by developing their English writing skills and ensuring they are academically prepared to enter a college program. Placement into DevEd or college-level courses is typically determined by standardized test scores, which influence not only students' educational trajectories but also carry economic consequences, as students with lower scores are required to complete one or two semesters of remedial coursework (Bickerstaff et al., 2022).

This study aims to improve English writing proficiency assessments and placement within U.S. community colleges, where DevEd support is most needed. Approximately 62% of individuals, ages 16 to 24, enrolled in colleges or universities in 2023, according to the United States Bureau of Labor Statistics[1]. At Tulsa Community College [2], where this research was conducted, around 30% of incoming students require developmental support in multiple areas, including English (reading and writing) and Math. The primary focus of this investigation is on English as an L1, though some observations may apply to the description of other languages.

Current automated placement tools, such as ACCUPLACER, align with academic standards but are not specifically calibrated to detect patterns unique to students requiring DevEd support. Moreover, ACCUPLACER's "black box" approach limits transparency and interpretability, posing challenges in educational contexts. This study seeks to address this issue by refining the linguistic descriptors used in placement assessments through the identification of key features that better capture writing proficiency in DevEd stu-

[a] https://orcid.org/0000-0001-8782-8377
[b] https://orcid.org/0000-0003-4603-4364

[1] https://www.bls.gov
[2] https://www.tulsacc.edu

Da Corte, M. and Baptista, J.
Refining English Writing Proficiency Assessment and Placement in Developmental Education Using NLP Tools and Machine Learning.
DOI: 10.5220/0013351500003932
In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 288-303
ISBN: 978-989-758-746-7; ISSN: 2184-5026

dents. By leveraging both human annotations and NLP-derived features, this research aims at enhancing placement accuracy and providing actionable insights for instructional support.

For automated placement systems to become effective tools for both faculty and students, the use of high-quality annotated corpora is essential, despite the time-intensive nature of annotation tasks. These annotations, crucial for training and testing automated scoring systems, enhance the reliability and applicability of placement decisions, ultimately supporting equitable access to education. This study addresses the need for accuracy in DevEd placements by refining and expanding an annotated dataset, integrating Natural Language Processing (NLP) tools for automatic feature extraction, and supplementing these with novel, manually encoded DevEd features.

The methodology followed employs a 300-essay corpus annotated with 11 DevEd-specific features, complemented by linguistic features automatically extracted through NLP tools, namely COH-METRIX[3] (McNamara et al., 2006) and the Common Text Analysis Platform (CTAP)[4] (Chen and Meurers, 2016). This corpus will be used to train and test various ML classifiers against both ACCUPLACER's automated classifications and human ratings, assessing classification accuracy to identify the most predictive features and the most effective algorithms for DevEd placements. A suite of well-known ML algorithms from the ORANGE[5] data-mining tool (Demšar et al., 2013) is used as the experiments with the models it makes available can be easily replicated, helping provide a framework for future comparisons.

Based on these motivations and the existing limitations mentioned, the purpose of this paper is to: (i) analyze linguistic features in an expanded corpus using COH-METRIX, CTAP, and a manually annotated set of DevEd-specific (DES) features to enhance text classification accuracy; (ii) assess the impact of these combined features on classification accuracy for student placement in DevEd courses through supervised ML experiments; (iii) compare classification accuracy rates achieved with the larger corpus and refined feature set to a previously established baseline; and (iv) offer actionable insights into students' linguistic abilities to support accurate placement and targeted instructional support in DevEd courses.

[3]https://141.225.61.35/CohMetrix2017/

[4]https://sifnos.sfs.uni-tuebingen.de/ctap/

[5]https://orangedatamining.com/

2 RELATED WORK

Concerns about the accuracy of automated classification systems like ACCUPLACER stem from the limited detail in the linguistic features considered during the classification process, as well as uncertainty around how these features are defined and ranked (Roscoe et al., 2014; Johnson et al., 2022). This lack of transparency makes it difficult to clearly explain to students which linguistic features were evaluated to determine their placement, what defines college-level writing, and which patterns they should address to improve their skills. Without such clarity, students may lack the necessary guidance to effectively enhance their writing abilities (Ma et al., 2023).

Studies suggest that up to one-third of placements may be inaccurate (Ganga and Mazzariello, 2019). Inaccurate placement in the context of DevEd often leads to students being assigned to courses that either underestimate or overestimate their writing abilities, both of which carry significant implications, not only economically, but for student outcomes and institutional effectiveness (Hughes and Li, 2019; Link and Koltovskaia, 2023; Edgecombe and Weiss, 2024).

Assessing students' linguistic skills solely through test scores may overlook critical aspects of their language abilities at the onset of their academic journey, highlighting the need for a more nuanced evaluation approach. As institutions seek to develop and support students' literacy-related skills, especially the descriptive quality of their written work (Kosiewicz et al., 2023), it becomes essential to explore and articulate the developmental aspects of writing for this population. Among the features worth exploring, the literature emphasizes the lexical and syntactic properties of a text, followed by the text's length, portraying them as key indicators of writing development (Stavans and Zadunaisky-Ehrlich, 2024).

Text structure, a key measure of text quality, constitutes another property worth exploring (Kyle et al., 2021; Feller et al., 2024). Texts that rely on single lines or simple, non-multilayered paragraphs are identified as needing remediation, while those with well-structured paragraphs—including a clear introduction, supporting statements, and a conclusion—are better aligned with academic writing standards (Da Corte and Baptista, 2025). Additionally, texts that demonstrate a varied grammatical repertoire, grammatical complexity, and purposeful grammatical choices are more in line with these standards, showcasing the writer's ability to use grammar effectively (Nygård and Hundal, 2024).

Studies on DevEd placement for English (L1) speakers have focused on refining texts ranking tasks

and improving classification accuracy (CA) through corpora annotation (Götz and Granger, 2024) and lexical and syntactic analysis supported by text mining (Lee and Lee, 2023). Key NLP tools, such as COH-METRIX (McNamara et al., 2006) and CTAP (Chen and Meurers, 2016)), have proven essential for assessing linguistic complexity across languages. While this study focuses on English, the identification and development of linguistic features for enhanced accuracy in automatic language proficiency classification is applicable beyond this language. For instance, (Leal et al., 2023) adapted COH-METRIX for Brazilian Portuguese and (Okinina et al., 2020) extended CTAP to Italian. (Akef et al., 2023) applied CTAP to Portuguese proficiency assessment, achieving 76% CA and emphasizing the role of feature selection. Similarly, (Wilkens et al., 2022) investigated lexical diversity and syntactic dependency relations for evaluating French language development.

Exploring further the above-mentioned contributions, lexical features such as *word n-grams*, *part-of-speech (POS) n-grams*, *POS-tag ratios*, and *type-token ratio* (TTR), which reflect the diversity of word types in a text, play a crucial role in these analyses, in addition to lexical variation features (e.g., noun, adjective, adverb, and verb variations), as well as metrics like lexical density, to capture the breadth of students' linguistic proficiency (Vajjala and Meurers, 2012). This confirms the importance of incorporating rich linguistic features into automated classification systems to enhance the accuracy of writing assessments across languages (Vajjala and Lučić, 2018; Vajjala, 2022).

By focusing on descriptive features that more accurately reflect native English proficiency, automatic classification systems have shown measurable improvements in placement accuracy (Da Corte and Baptista, 2024c; Da Corte and Baptista, 2024b; Da Corte and Baptista, 2025), providing valuable insights into students' readiness for college and their linguistic progression over time. Studies such as (Pal and Pal, 2013) have demonstrated this potential by using the WEKA machine-learning platform with models like Naïve Bayes, Multilayer Perceptron, and Decision Trees, achieving a classification accuracy (CA) close to 90% in course placement. Similarly, (Filighera et al., 2019) utilized Neural Networks and embeddings to categorize texts by reading level with approximately 80% accuracy, while (Hirokawa, 2018) and (Duch et al., 2024) further demonstrated the effectiveness of classical ML algorithms—including Naïve Bayes, Neural Networks, and Random Forest—in feature selection and outcome prediction, with CA rates above 90%.

In alignment with the motivation and objectives of this study, the findings presented in the state-of-the-art review confirm the potential of machine learning models to enhance DevEd placement by emphasizing relevant linguistic features, helping institutions better support students both academically and linguistically (Bickerstaff et al., 2022; Giordano et al., 2024).

3 METHODS

The methodology for this study involves a systematic and detailed setup that prioritizes ethical participant selection and focuses on linguistic descriptor reliability for DevEd placement, all done in accordance with the Institution's Review Board (IRB) protocols (ID #22-05[6]). The methodology was designed to adhere to guidelines sensitive to the unique academic challenges experienced by this group.

3.1 Corpus

This study extends a previous classification task conducted with a small corpus of 100 text samples (Da Corte and Baptista, 2024a; Da Corte and Baptista, 2025) by adding 200 more essays, resulting in a systematically selected corpus of 300 essays totaling 97,339 tokens[7]. The corpus focuses on non-college-level classifications to enhance the precision of placement for students requiring DevEd support. The texts were written by students seeking college admission during the 2021-2022 and 2023-2024 academic years and were drawn from a larger pool of 1,000 essays within the institution's standardized entrance exam database. The selected essays cover 11 distinct writing topics, including *Is History Valuable*, *Necessary to Make Mistakes*, *Acquisition of Money*, *Differences Among People*, *Happiness Not an Accident*, and *Independent Ideas*, to mention a few. These essays were written in a controlled, proctored environment at the institution's testing center, ensuring no Internet access or editing tools were available.

The selection criteria aimed at compiling a corpus focused on non-college-level classifications, accurately representing students in DevEd and their placement levels as determined by ACCUPLACER into Levels 1 or 2. The corpus was balanced by level, with essays averaging 324 tokens each, making it well-suited for examining classification by level. This dataset provides a valuable foundation for assessing

[6]https://www.tulsacc.edu/irb

[7]Corpus dataset to be made available after paper publication.

language proficiency among community college students. Demographic information—including gender, first language, years of English studied in high school, and race—was available for 67% of participants but was omitted from this stage of analysis. This data will be analyzed in detail in a future study.

3.2 Corpus Annotation

As an initial step in this study, two trained raters annotated the corpus with a refined set of 11 DES features devised by (Da Corte and Baptista, 2024a)[8]. Selected through an open call for volunteers, the annotators were equally represented in terms of gender distribution (1 male and 1 female), were both native English speakers with advanced English skills, held at least a Bachelor's degree, and had experience in higher education.

The refined set of features derived from an initial list of 21 and was narrowed down to 11 using ORANGE's feature selection tool and the Information Gain ranking method. These features, identified as the most discriminative in previous experiments (Da Corte and Baptista, 2022; Da Corte and Baptista, 2024a), are summarized in Table 1. They are grouped into 4 (pattern) clusters: Orthographic (ORT), Grammatical (GRAMM), Lexical and Semantic (LEXSEM), and Discursive (DISC), each reflecting key aspects of foundational language proficiency. While most features are classified as negative (-), representing deviations from proficiency and identifying errors, others are positive (+), serving as indicators of proficiency and advanced language use.

ORT patterns capture basic language structure through grapheme alterations, punctuation, and the use of contractions (generally avoided in academic writing unless otherwise specified). GRAMM patterns measure text quality through verb agreement, referential pronoun usage, and the omission or addition of a part of speech, which often interferes with the meaning of a statement. LEXSEM patterns include the use of multiword expressions (MWE) and lexical accuracy. Lastly, DISC patterns highlight writers' ability to extend discourse with reasoned arguments and examples. These features were integrated with the previously mentioned NLP-derived features, automatically extracted from COH-METRIX (106) and CTAP platforms (328)[9], and used in the supervised ML experiments detailed in Section 4.

For the annotation of these features on the texts, Labelbox[10] was selected as the web annotation tool to use. Labelbox has been tested before and identified as a tool that simplifies the data labeling process in a way that can be used to train Artificial Intelligence (AI) models, and enables the creation of high-quality annotated datasets (Colucci Cante et al., 2024).

3.3 Annotation Task Assessment

After manually annotating the corpus, all tags were meticulously tracked and processed through Labelbox. Krippendorff's Alpha (K-alpha) inter-rater reliability coefficient was computed to evaluate annotation reliability. Given the size of the dataset, these calculations were performed via Excel formulas. For the interpretation of K-alpha scores, the following agreement thresholds and interpretations, set forth by (Fleiss and Cohen, 1973), were followed:

below 0.20 - *slight* (Sl);
between 0.21 and 0.39 - *fair* (F);
between 0.40 and 0.59 - *moderate* (M);
between 0.60 and 0.79 - *substantial* (Sb);
above 0.80 - *almost perfect* (P);

The two annotators agreed on the assessment and tagging of 79,805 tokens and disagreed on 17,534 tokens. The observed proportion of agreement (P_o) was 0.819, while the expected agreement by chance (p_e) was 0.5 (or 50%), given two possible outcomes with equal probability. Using the formula:

$$K - alpha = \frac{P_o - 0.5}{1 - 0.5}$$

the inter-rater reliability score was calculated as $k = 0.640$, indicating "substantial agreement". Considering the complexity of the task, this level was deemed adequate. The reasons for disagreeing on the 17,534 tokens will be analyzed in a future study.

A third annotator, with a similar background to the other two, was engaged in the task assessment to facilitate consensus and establish a gold standard (or reference) annotation. This process resulted in a final set of 17,342 tags applied and used for analysis. Table 2 provides a detailed count of tagged features per level, maintaining the order of features as presented in Table 1. The table also highlights the uneven distribution of tags, with 38% found in texts classified by ACCUPLACER as Level 1 and 62% as Level 2. Notably, this feature distribution offers foundational insights into the linguistic attributes that typically distinguish Levels 1 and 2 texts, serving as an initial prototype for defining the unique characteristics of each level.

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

[9]The definitions of COH-METRIX and CTAP are well-documented in the literature and adopted as cited.

[10]https://docs.labelbox.com/

Table 1: DevEd-specific (DES) features summary.

Patterns	Description	Features Clustered
Orthographic (ORT)	Patterns representing the foundational language skills needed to represent words and phrases.	(-) Grapheme (addition, omission, transposition, and capitalization) (ORT) (-) Word split (WORDSPLIT) (-) Punctuation used & Contractions (PUNCT)
Grammatical (GRAMM)	Patterns evidencing the quality of text production.	(-) Word omitted (WORDOMIT) (-) Word repetition (WORDREPT) (-) Verb agreement (VAGREE) (-) Pronoun-alternation referential (ALTERN)
Lexical & Semantic (LEXSEM)	Patterns contributing to the structuring of a writer's discourse.	(-) Word precision (PRECISION) (+) Multiword expressions (MWE)
Discursive (DISC)	Patterns exhibiting the writer's ability to produce— extended discourse.	(+) Argumentation with reason (REASON) (+) Argumentation with example (EXAMPLE)

Table 2: Distribution of DevEd-specific (DES) features across the corpus.

Polarity	Feature	Level 1	%	Level 2	%	Dif (L1-L2)	Result
-	ORT	2,528	38.18%	2,504	23.36%	-14.83%	Improved
-	WORDSPLIT	95	1.43%	72	0.67%	-0.76%	Almost no improvement
-	PUNCT	1,390	20.99%	2,082	19.42%	-1.57%	Slightly improved
-	WORDOMIT	435	6.57%	462	4.31%	-2.26%	Slightly improved
-	WORDREPT	71	1.07%	98	0.91%	-0.16%	Almost no improvement
-	VAGREE	178	2.69%	106	0.99%	-1.70%	Slightly improved
-	ALTERN	37	0.56%	46	0.43%	-0.13%	Almost no improvement
-	PRECISION	599	9.05%	745	6.95%	-2.10%	Slightly improved
+	MWE	1,038	15.68%	4,101	38.25%	22.57%	Improved
+	REASON	200	3.02%	356	3.32%	0.30%	Almost no improvement
+	EXAMPLE	50	0.76%	149	1.39%	0.63%	Almost no improvement
Total	-	**6,621**	-	**10,721**	-	-	-

Overall, Level 1 texts exhibited more frequent instances of foundational issues (PUNCT and ORT) and demonstrated less cohesion and complexity. In contrast, Level 2 texts displayed fewer lower-order errors and greater exemplification of higher-order features, such as MWE. As writing develops, texts tend to incorporate more sophisticated features with corrections focusing on fine-tuning elements that enhance clarity and refine expression (Nygård and Hundal, 2024). As an illustration, two sample texts representing Levels 1 and 2 are provided in the Appendix.

The primary improvements noted between texts in Levels 1 and 2 were a sharp reduction in ORT errors (-14.83%) and a notable increase in the use of multiword expressions (MWE) (+22.57%), corroborating previous research findings on the connection between MWE use and higher proficiency levels (Laporte, 2018; Kochmar et al., 2020; Pasquer et al., 2020). Additionally, a slight improvement was observed in the reduced incidence of several negative features, with proficiency levels improving by approximately 1.6% to 2.3% for these features. The distribution of the remaining features shows no significant differences, with changes below 1%.

Building on this tagging framework, the ranking of DES features was compared with the original subcorpus of 100 text samples, 200 texts, and the full corpus, 300 text samples, as shown in Table 3. The ranking was determined using the Information Gain scoring method available in ORANGE, selected due to its superior classification accuracy (CA) demonstrated in a prior study (Da Corte and Baptista, 2024c). Consequently, this method was applied to all classification tasks described in Section 4.

The ranking of DES features for the 100- and the 200-text samples yielded a "very strong" Spearman correlation score[11] (Schober et al., 2018) of $\rho = 0.927$ with a two-tailed *p*-value of 0, indicating consistency in feature ranking across studies despite observable differences. This correlation, however, does not preclude some shifts in the relative importance of certain features, likely influenced by differences in corpus size. Note that the content of each sample does not overlap. Observable changes include higher rankings for MWE, EXAMPLE (argumentation of key concepts; support of one's position with examples.), and VAGREE (lack of agreement between the subject of the sentence and the conjugated form of the verb) in the 200 text samples, while the PRECISION (imprecise use of words attending to its meaning in the sentence) feature dropped significantly from 4th to 11th place, in this sub-corpus.

The final ranking with all 300 text samples combined portrays VAGREE, EXAMPLE, and ORT (denoting orthographic errors) as the top 3 features with comparable Information Gain scores. Comparison of the Spearman ranking coefficient is also "very high", with $\rho = 0.948$ for the 100 samples vs. 300 samples corpus ($<$ overlap), and $\rho = 0.977$ for the 200 vs. 300 samples ($>$ overlap). This indicates a consistent ranking of the features across the subcorpora.

[11] https://www.socscistatistics.com/tests/spearman/

Table 3: Comparison of DES Feature Rankings Using Information Gain Scores for 100, 200, and 300 Texts.

100 Texts			200 Texts			300 Texts		
Rank	**Feature**	**Info. Gain**	**Rank**	**Features**	**Info. Gain**	**Rank**	**Features**	**Info. Gain**
1	ORT	0.084	1	MWE	0.260	1	VAGREE	0.131
2	EXAMPLE	0.080	2	EXAMPLE	0.171	2	EXAMPLE	0.124
3	WORDOMIT	0.080	3	ORT	0.140	3	ORT	0.112
4	PRECISION	0.072	4	WORDOMIT	0.126	4	MWE	0.088
5	WORDREPT	0.062	5	VAGREE	0.126	5	WORDOMIT	0.075
6	VAGREE	0.056	6	ALTERN	0.081	6	REASON	0.045
7	MWE	0.050	7	WORDREPT	0.078	7	WORDREPT	0.039
8	REASON	0.039	8	REASON	0.069	8	ALTERN	0.034
9	ALTERN	0.038	9	WORDSPLIT	0.060	9	PRECISION	0.014
10	PUNCT	0.037	10	PUNCT	0.034	10	WORDSPLIT	0.012
11	WORDSPLIT	0.036	11	PRECISION	0.004	11	PUNCT	0.011

3.4 Skill-Level Classification and Assessment

In addition to assessing the DevEd-specific features marked in the corpus, all 300 full-text samples were classified by the same two annotators according to skill level, using DevEd-level definitions adapted from the Institution's course descriptions:

> **DevEd Level 1:** if essays demonstrated a need for improvement in general English usage, including grammar, spelling, punctuation, and the structure of sentences and paragraphs;
>
> **DevEd Level 2:** if essays required targeted support in specific aspects of English, such as sentence structure, punctuation, editing, and revising; or
>
> **College Level:** if essays were written accurately and displayed appropriate English usage at the college academic level.

The two evaluators agreed on the skill level assigned to 260 text samples, while they differed on 40. Among the agreed-upon samples, 122 were classified as Level 1, 134 as Level 2, and 4 as College level. To measure the level of agreement, the Krippendorff's Alpha (K-alpha) inter-rater reliability coefficient was calculated using the ReCal-OIR[12] tool (Freelon, 2013) for ordinal data.

According to (Fleiss and Cohen, 1973), a "moderate" level of agreement 0.473 was obtained. While this score indicates some consistency in the ratings between the annotators, it also highlights the complexity involved in assessing proficiency skills. This suggests that incorporating additional descriptive linguistic criteria, such as numerical subscales to quantify misspelled words, beyond the outlined DevEd level descriptions, may enhance automatic classification accuracy.

Given the moderate agreement score, further analysis was necessary to resolve discrepancies in classifications. For the 40 cases where differences between human classifications occurred, the average of the levels assigned for each text was calculated, and a final classification was attributed. This process resulted in 14 cases being classified as Level 1, 22 as Level 2, and 4 as College level.

With the discrepancies finally resolved, the DevEd corpus now comprised 136 essays classified as Level 1, 156 as Level 2, and 8 as College-level. To maintain balance between the levels, random resampling was conducted for Level 2 to reduce the number of essays to 136, equal to Level 1. The College-level essays were discarded at this stage, as this study focuses on a two-level DevEd classification task.

This human-rated assessment adds a level of depth to the analysis, as annotators, unlike automated systems like ACCUPLACER, can grasp the subtleties of language use and the complexities involved in student writing assessment.

3.4.1 Comparing ACCUPLACER vs. Human Assessment

The classification performance of the ACCUPLACER system with human annotators is compared in this section, providing insights into alignment and discrepancies across subcorpora of 100, 200, and all 300 texts, using Pearson coefficient (Cohen, 1988).

The correlation between ACCUPLACER and human classifications is summarized in Table 4.

Table 4: Pearson Correlation Comparison: Human Raters vs. Accuplacer Across Subcorpora.

Subcorpus	Pearson
100	0.242
200	0.366
300	0.305

The Pearson coefficients consistently indicated "weak" correlation scores for 100, 200, and for the full corpus. Similarly, the K-alpha coefficient, computed for all 300 texts, yielded a "fair" score of $k = 0.312$ confirming ACCUPLACER'S limitations in aligning with human raters, particularly in identifying

[12] https://dfreelon.org/utils/recalfront/recal-oir/

nuanced linguistic features and transitions between proficiency levels.

To further illustrate the discrepancies between ACCUPLACER and human classifications, confusion matrices (Tables 5,6, 7) provide a detailed breakdown of specific instances where ACCUPLACER encounters challenges in accurately predicting proficiency levels.

Table 5 shows that ACCUPLACER aligned with human classifications for 79 texts (accuracy, 79%). All 21 errors occurred when Level 1 texts were misclassified as Level 2. The results suggest that this automatic classification system has some limitations when distinguishing Level 1 from Level 2 and the need for improved recognition of Level 1 linguistic features.

Table 5: Accuplacer vs. Human Assessment: 100 Texts.

	Accuplacer L1	Accuplacer L2	Sum
Human L1	50	21	71
Human L2	0	29	29
Sum	50	50	100

With the subcorpus of 200 text samples, as per Table 6, ACCUPLACER aligned with the human classification for 157 texts (accuracy, 78.5%), a similar ratio as with the subcorpora of 100 texts. However, the misclassification of 35 Level 2 texts as Level 1 significantly impacts overall accuracy. Additionally, 8 College-level texts were not recognized by Accuplacer as proficient texts, despite humans identifying linguistic features corresponding to advanced, college-level writing. These 8 texts were excluded from both this matrix and the subsequent one.

Table 6: Accuplacer vs. Human Assessment: 200 Texts.

	Accuplacer L1	Accuplacer L2	Sum
Human L1	65	0	65
Human L2	35	92	127
Sum	100	92	192

For the 300 text samples, ACCUPLACER aligned with the human classification for 236 texts (accuracy, 79%). In Table 7, a misclassification of 56 texts (35 Level 2 texts misclassified as Level 1; 21 Level 1 texts misclassified as Level 2) was noted, indicating ACCUPLACER'S difficulty in distinguishing these levels. Furthermore, ACCUPLACER'S missed all 8 College-Level texts classified by humans, suggesting that improvements are needed to better identify linguistic features that signal higher proficiency and transitions between Levels 1 and 2.

Table 7: Accuplacer vs. Human Assessment: 300 Texts.

	Accuplacer L1	Accuplacer L2	Sum
Human L1	115	21	136
Human L2	35	121	156
Sum	150	142	292

These misclassifications have broader implications beyond accuracy scores. Misplacing approximately 1 in 5 students (approximately 20%) at a level above or below their true skill carries significant pedagogical consequences. These include not only extra costs in time and money for both students and institutions but also potential impacts on learning outcomes and student success.

4 MACHINE LEARNING FOR DevEd PLACEMENT

As mentioned before, this study builds on earlier work (Da Corte and Baptista, 2024c) by integrating 445 features from three sources to classify text samples by DevEd level: 106 from COH-METRIX, 328 from CTAP, and the refined set of 11 DES features manually annotated on the corpus. Supervised ML methods now use this data in a set of experiments aimed at (i) identifying the most relevant linguistic features for classification and (ii) determining the ML algorithm achieving the highest classification accuracy (CA).

While this represents a classical ML approach to the classification problem at hand, future research will expand on these findings by incorporating pre-trained Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPT), to assess their ability to align textual features with refined descriptors and generate interpretative explanations for classification outcomes.

These experiments were set in two scenarios: (i) using the full corpus with classification levels automatically assigned by ACCUPLACER; and (ii) using the same corpus with classification levels assigned by human annotators. In each scenario, experiments were first conducted with all 445 features, followed by additional experiments where features were added incrementally in bundles of 10 until classification results reached asymptotic results.

The data-mining tool ORANGE was selected for analysis and modeling for its usability and the diversity of ML tools and algorithms it makes available. A total of 10 well-known ML algorithms were selected from the set available in ORANGE (listed alphabetically): (i) *Adaptive Boosting* (AdaBoost); (ii) *CN2 Rule Induction* (CN2); (iii) *Decision Tree* (DT) (iv) *Gradient Boosting* (GB); (v) *k-Nearest Neighbors* (kNN); (vi) *Logistic Regression* (LR); (vii) *Naïve Bayes* (NB); (viii) *Neural Network* (NN); (ix) *Random Forest* (RF); and (x) *Support Vector Machine* (SVM). The default configuration of these learners was selected.

Figure 1 shows the basic workflow adopted for this study.

To train and evaluate the models, the Orange TEST&SCORE widget was employed. Classification Accuracy (CA) was the primary evaluation metric, aligned with this study's goals. In cases where CA values were identical, Precision (Prec) was used as a secondary criterion to rank the models. Given the corpus size, a 3-fold cross-validation was implemented using the DATA SAMPLER widget, allocating 2/3 of the corpus for training and 1/3 for testing. Additionally, the RANK widget facilitated an evaluation of each feature's discriminative power for the task, while a Confusion Matrix provided insights into detailed breakdown of the classification results.

Previously, a baseline has been established, setting a CA of 0.727 using the Random Forest (RF) algorithm (Da Corte and Baptista, 2024c). This relatively high baseline reflects RF's recognized effectiveness in machine learning applications to writing assessment (Huang, 2023). Although the classification accuracy is high, the baseline represents a concerning value. A CA of 0.727 translates to approximately 3 out of 10 students being misclassified in DevEd following the ACCUPLACER writing assessment. This misclassification rate raises concerns about the methodology of current automatic placement systems, with potentially far-reaching implications for student success and the allocation of institutional resources. Addressing these challenges aligns with the objectives of this study, which seeks to enhance placement accuracy.

4.1 Scenario 1: Classification Using ACCUPLACER-Assigned Skill Levels

For this classification task, text sample units were classified using all 445 features with the skill-level classifications automatically assigned by ACCUPLACER as the target variable (150 texts Level 1; 150 texts Level 2). The results of this experiment are presented in Table 8.

Table 8: CA scores, F1, Precision, and Recall for all models with 445 features and ACCUPLACER's classifications.

Model	CA	F1	Prec	Recall
CN2	0.655	0.653	0.655	0.655
AdaBoost	0.675	0.675	0.675	0.675
DT	0.730	0.730	0.731	0.730
RF	**0.750**	0.748	0.753	0.750
kNN	**0.755**	0.751	0.764	0.755
NN	**0.770**	0.768	0.773	0.770
LR	**0.780**	0.780	0.780	0.780
NB	**0.785**	0.782	0.794	0.785
SVM	**0.795**	0.794	0.796	0.795
GB	**0.800**	0.798	0.806	0.800

When all features were combined, 7 out of the 10 models tested achieved higher CA scores, surpassing the initial baseline of 72.7% - from 75% with RF up to 80% with GB (highlighted in bold). This indicates that the combined features hold significant potential for improving student placement and ensuring they receive the necessary support as they begin college. With this enhanced feature set, at least 8 out of 10 students would now be properly placed, compared to the previous ratio of 7 out of 10.

To address the question of which features contribute most to improved placement accuracy, the most discriminative features from Coh-Metrix, CTAP, and DES were identified using the Information Gain ranking method. Table 9 provides the topmost 30.

The top-ranked features, overall, were mostly from CTAP (99%), followed by 2 COH-METRIX features, *Word count*, *Number of words* ranking 35th (Information Gain: 0.242) and *Sentence count*, *number of sentences* ranking 88th (Information Gain: 0.162) and one DES feature, VAGREE ranking 117th (Information Gain: 0.131). The next DES feature ranked was in 121st place, EXAMPLE (Information Gain: 0.124). Notably, VAGREE had previously ranked 7th when the experiments were conducted with a smaller text sample unit size of 100 (Da Corte and Baptista, 2024c).

All features were then grouped into sets of 10, and bundles were used in several ML experiments to identify optimal CA scores performance before results reached an asymptote. The CA scores for the top 120 selected features for all 10 ML algorithms are detailed in Table 10. CA scores in bold indicate the highest score achieved with different feature combinations, while scores exceeding the previously mentioned baseline of 0.727 (in Section 4) are italicized for each model and bundle. Figure 2 depicts the outcomes of Experiment 4.1 for a more in-depth evaluation of the results obtained.

Most models consistently outperformed the previous baseline of 0.727 CA, except for the AdaBoost and CN2 algorithms, which fell below this value in most instances. The NN model quickly peaked in the first run, achieving a CA score of 0.800 with only 10 combined features. While its performance slightly declined in subsequent experiments, it consistently remained above 0.727. Similarly, the NB model demonstrated strong performance with consistent CA scores of 0.770 or higher, reaching asymptote results after a total of 30 combined features. This consistent behavior is likely due to NB's known great computing efficiency and adaptability to text classification tasks (Pajila et al., 2023).

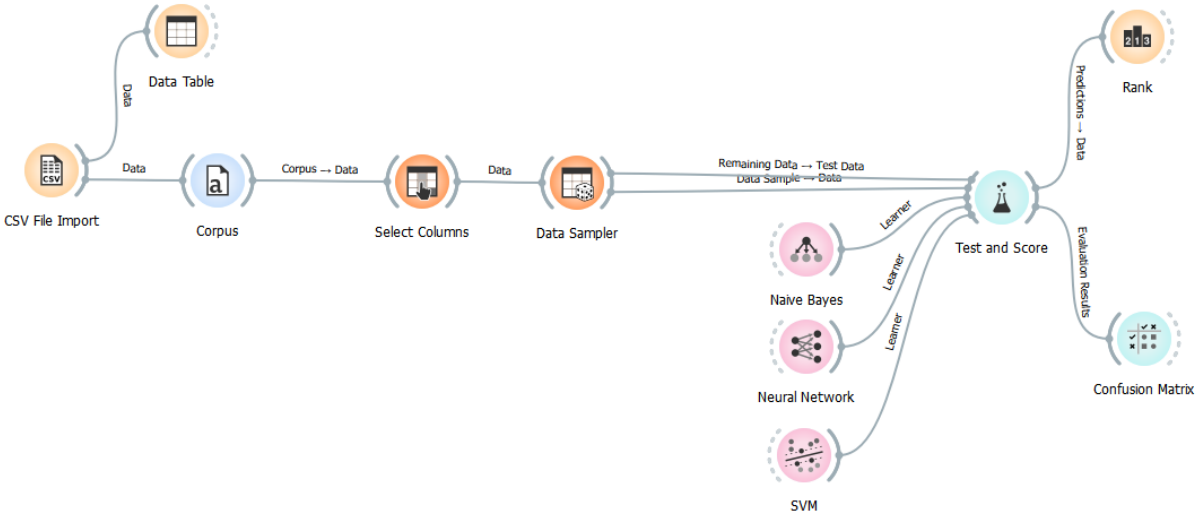


Figure 1: ORANGE workflow setup. The NB, NN, and SVM algorithms are displayed merely as representatives of the chosen learners.

Table 9: Scenario 1: Combined top-ranked 30 features from COH-METRIX, CTAP, and DES by Information Gain scores.

Rank	Source	Feature Description	Info. gain
1	CTAP	Lexical Sophistication: Easy word types (NGSL)	0.341
2	CTAP	Number of Tokens with More Than 2 Syllables	0.333
3	CTAP	Number of Word Types (including Punctuation and Numbers)	0.331
4	CTAP	Number of Word Types (excluding Punctuation and numbers)	0.327
5	CTAP	Number of Word Types	0.327
6	CTAP	Number of POS Feature: Lexical word Types	0.326
7	CTAP	Number Of Letters	0.311
8	CTAP	Number of Word Types with More Than 2 Syllables	0.303
9	CTAP	Number of POS Feature: Verb Types (including Modals)	0.296
10	CTAP	Number of POS Feature: Adjective and Adverb Types	0.288
11	CTAP	Lexical Sophistication: Easy verb types (NGSL)	0.275
12	CTAP	Number of POS Feature: Preposition Types	0.273
13	CTAP	Number of syllables	0.272
14	CTAP	Lexical Sophistication: Easy lexical types (NGSL)	0.271
15	CTAP	Number of POS Feature: Verb Types (without Modals)	0.271
16	CTAP	Lexical Sophistication Feature: SUBTLEX Word Frequency per Million (AW Type)	0.270
17	CTAP	Lexical Sophistication: Easy lexical tokens (NGSL)	0.269
18	CTAP	Number of POS Feature: Lexical word Lemma Types	0.269
19	CTAP	Number of POS Feature: Noun Types	0.268
20	CTAP	Number of POS Feature: Noun Lemma Types	0.266
21	CTAP	Number of POS Feature: Verb Lemma Types	0.265
22	CTAP	Number of POS Feature: Noun Tokens	0.263
23	CTAP	Number of Syntactic Constituents: Complex Noun Phrase	0.263
24	CTAP	Number of POS Feature: Lexical word Tokens	0.262
25	CTAP	Number of POS Feature: Adverb in base form Types	0.262
26	CTAP	Number of POS Feature: Adverb Types	0.260
27	CTAP	Number of POS Feature: Verb Tokens (without Modals)	0.258
28	CTAP	Number of Syntactic Constituents: Declarative Clauses	0.253
29	CTAP	Lexical Sophistication: Easy verb tokens (NGSL)	0.251
30	CTAP	Number of Unique Words	0.250

Table 10: Scenario 1: CA scores for top 120 features in bundles of 10 across 10 ML models.

	Classification Accuracy (CA) Scores											
Model	**10ft**	**20ft**	**30ft**	**40ft**	**50ft**	**60ft**	**70ft**	**80ft**	**90ft**	**100ft**	**110ft**	**120ft**
DT	0.700	*0.745*	0.665	0.655	0.660	0.660	*0.735*	*0.735*	*0.770*	*0.760*	*0.755*	0.725
AdaBoost	0.715	0.695	0.640	0.665	0.720	*0.730*	0.715	0.715	*0.765*	0.695	0.710	0.705
CN2	0.715	0.680	0.660	0.715	0.695	0.715	0.695	0.695	0.700	0.695	0.695	0.695
GB	*0.755*	*0.745*	*0.735*	0.705	0.710	*0.745*	*0.765*	*0.760*	*0.765*	*0.760*	***0.780***	*0.775*
NB	*0.770*	*0.790*	***0.790***	*0.785*	***0.785***	***0.780***	***0.780***	***0.785***	***0.785***	***0.780***	***0.780***	***0.780***
RF	*0.770*	*0.745*	0.720	*0.740*	*0.760*	*0.750*	*0.730*	*0.775*	*0.765*	*0.740*	*0.745*	*0.755*
LR	*0.770*	*0.780*	*0.780*	0.705	*0.755*	0.700	0.710	0.700	0.710	0.700	0.715	*0.755*
kNN	*0.775*	*0.740*	*0.735*	*0.765*	*0.770*	*0.760*	*0.745*	*0.735*	*0.775*	*0.770*	*0.770*	*0.770*
SVM	*0.790*	***0.795***	*0.780*	*0.780*	*0.765*	*0.770*	*0.770*	*0.775*	*0.770*	*0.770*	*0.770*	*0.775*
NN	***0.800***	*0.790*	*0.760*	***0.790***	*0.760*	*0.760*	*0.755*	*0.765*	*0.750*	*0.760*	*0.760*	*0.740*

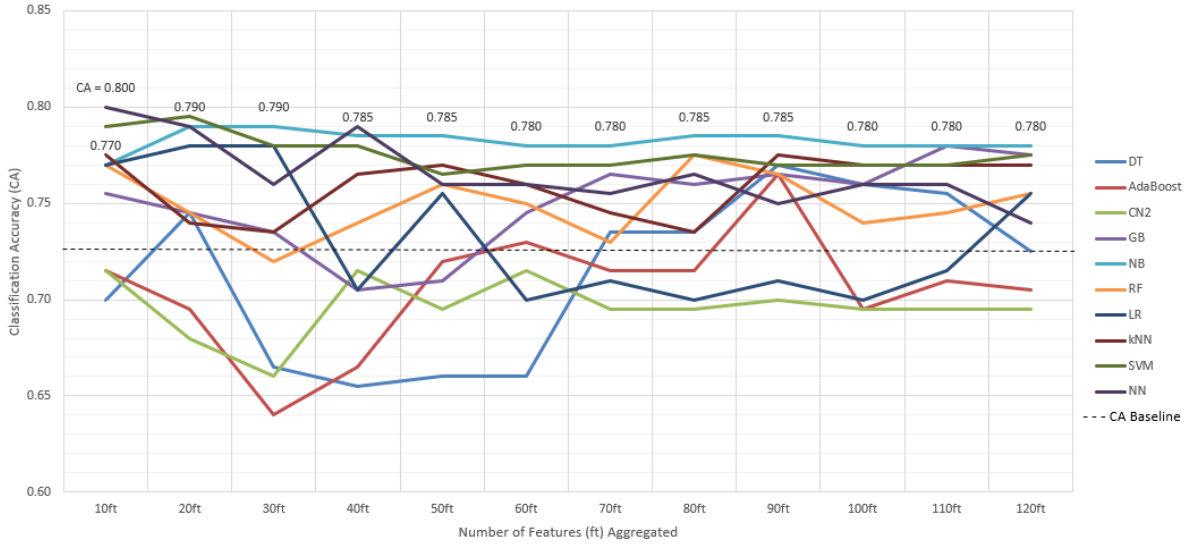


Figure 2: Scenario 1: Machine-learning algorithms performance based on feature aggregation with Accuplacer classification.

Assessing the Impact of DES Features

To better gauge the impact of DES features on the overall classification task, the 30 CTAP features—where NB reached its asymptote—were combined with all 11 DES features. The resulting CA scores are shown in Table 11, with scores exceeding NB's CA of 0.790 highlighted in bold. The final column compares the CA differences with 30 CTAP features alone to their combination with DES features, where positive values indicate improved accuracy, and negative values reveal a hindrance to the models' performance by this addition.

Table 11: CA scores: top-ranked 30 plus DES features.

CA Model Performance Comparison			
Model	**w/30ft**	**w/30ft+11 DES**	**Difference**
DT	0.665	0.750	+8.5%
AdaBoost	0.640	0.705	+6.5%
CN2	0.660	0.640	-2.0%
GB	0.735	0.760	+2.5%
NB	**0.790**	**0.795**	**+0.5%**
RF	0.720	0.755	+3.5%
LR	0.780	0.785	+0.5%
kNN	0.735	0.735	No difference
SVM	0.780	**0.810**	**+3.0%**
NN	0.760	**0.805**	**+4.5%**

Most models improved their performance with the addition of DES features. DT and AdaBoost showed the most significant gains (+8.5% and +6.5%), while CN2 experienced a slight decline in accuracy (-2%). The NB remained stable with a modest 0.5% gain. SVM and NN achieved CA scores of 0.810 and 0.805, improving by 3% and 4.5%, respectively. These findings highlight the significance of DES features as key indicators of students' writing abilities in DevEd.

4.2 Scenario 2: Classification Using Human-Assigned Skill Levels

Lastly, the resampled dataset of 272 text units (136 for Level 1 and 136 for Level 2), balanced by level following human assessment as described in Section 3.4, was classified using all 445 features, with the human-assigned skill levels as the target variable. The results from this experiment are presented in Table 12.

Table 12: CA scores, F1, Precision, and Recall for all models with 445 features with Human classifications.

Model	**CA**	**F1**	**Prec**	**Recall**
CN2	0.552	0.553	0.554	0.552
LR	0.558	0.555	0.564	0.558
AdaBoost	0.608	0.608	0.609	0.608
kNN	0.613	0.613	0.615	0.613
NB	0.635	0.631	0.649	0.635
GB	0.635	0.635	0.639	0.635
NN	0.657	0.657	0.662	0.657
DT	0.674	0.674	0.674	0.674
RF	0.674	0.674	0.674	0.674
SVM	**0.713**	0.711	0.715	0.713

Overall, the ML algorithms did not surpass the previously established 0.727 baseline CA score. However, the SVM model was the highest-performing model, with a score of 0.713. To further understand the algorithms' performance, the most discriminative attributes across COH-METRIX, CTAP, and DES were identified using the Information Gain ranking method. The top 30 features are included in Table 13.

When compared to the ranking of the features presented in Section 4.1, some shifts in feature rankings were noted, reflecting differences in how human raters and automated systems interpret linguis-

Table 13: Scenario 2: Combined top-ranked 30 features from COH-METRIX, CTAP, and DES by Information Gain scores.

Rank	Source	Feature Description	Info. gain
1	COH-Metrix	Paragraph length, number of sentences in a paragraph, mean	0.144
2	COH-Metrix	Lexical diversity, VOCD, all words	0.142
3	CTAP	Lexical Sophistication: Easy noun tokens (NGSL)	0.124
4	COH-Metrix	LSA given/new, sentences, mean	0.124
5	CTAP	Lexical Richness: Type Token Ratio (STTR NGSLeasy Nouns)	0.118
6	COH-Metrix	Lexical diversity, MTLD, all words	0.115
7	CTAP	Number of POS Feature: Plural noun Types	0.109
8	COH-Metrix	LSA given/new, sentences, standard deviation	0.109
9	**DES**	**EXAMPLE**	0.103
10	CTAP	POS Density Feature: Existential There	0.102
11	COH-Metrix	Positive connectives incidence	0.100
12	COH-Metrix	LSA overlap, adjacent sentences, mean	0.100
13	CTAP	Number of POS Feature: Existential there Types	0.099
14	CTAP	Number of POS Feature: Preposition Types	0.099
15	COH-Metrix	WordNet verb overlap	0.098
16	CTAP	Number of Syntactic Constituents: Postnominal Noun Modifier	0.097
17	CTAP	Number of Word Types (including Punctuation and Numbers)	0.097
18	CTAP	Lexical Sophistication Feature: SUBTLEX Logarithmic Word Frequency (AW Type)	0.096
19	CTAP	Number of POS Feature: Existential there Tokens	0.096
20	CTAP	Lexical Sophistication: Easy noun types (NGSL)	0.094
21	CTAP	Number of POS Feature: Verbs in past participle form Types	0.092
22	COH-Metrix	LSA verb overlap	0.092
23	CTAP	Number of Unique Words	0.092
24	CTAP	Number of Tokens with More Than 2 Syllables	0.090
25	CTAP	Number of Word Types with More Than 2 Syllables	0.086
26	CTAP	Lexical Richness: HDD (excluding punctuation and numbers)	0.086
27	CTAP	POS Density Feature: Possessive Ending	0.086
28	CTAP	Number of Syntactic Constituents: Complex Noun Phrase	0.084
29	CTAP	Number of POS Feature: Plural noun Tokens	0.083
30	CTAP	Referential Cohesion: Global Lemma Overlap	0.083

tic patterns. Most of the highest-ranked features derived from CTAP (67.5%), with a much smaller proportion from COH-METRIX (11%). Two DES features (5%)—EXAMPLE(Information Gain: 0.103) and ORT (Information Gain: 0.082)—were ranked 9th and 31st, respectively.

The presence of these two DES features (EXAMPLE and ORT) within the top 31 may highlight the importance of human-derived features in the task, though a good approximation of the other features might have been obtained by mechanical methods as well. These features also seem to enable a more detailed assessment of students' writing abilities before entering an academic program, providing valuable insights into their communication effectiveness (Kim et al., 2017).

To prepare for the next step in this classification task, all features were also tested by grouping them in bundles of 10, using their Information Gain scores, and incrementally adding them to observe asymptotic trends in the results. Table 14 summarizes the impact of these feature bundles (10 bundles, 120 features total) on the classification accuracy (CA) of the ML algorithms used. Bolded CA scores represent the highest score achieved by the ML algorithms with different feature bundles, while scores exceeding the baseline of 0.727 are italicized. Figure 3 depicts the outcomes of Experiment 4.2.

Out of all the algorithms tested, four (CN2, DT, AdaBoost, and kNN) underperformed relative to the 0.727 baseline, with CN2 and kNN yielding the lowest results and showing no significant CA improvements as features were added. Conversely, SVM, RF, LR, and NB performed better, often surpassing the 0.727 baseline. Notably, the NB model achieved a CA of 0.779 with just 10 features, though scores declined slightly as more features were added, remaining above 0.727 with the first 30 features. LR peaked at 0.785 with 40 features, an improvement of nearly 6% from the baseline, before its accuracy decreased, on average, by about 12% in subsequent iterations. SVM maintained consistent performance above 0.727, in most instances, reaching an asymptotic line at 70 features (CA = 0.740).

The enhanced CA score of almost 80% achieved by LR indicates some improvements by signaling that instead of misclassifying 3 students, only 2 face incorrect placement, with one more student now receiving the essential support needed in college.

Assessing the Impact of DES Features

After achieving a peak CA of 0.785 with LR with 40 features, a combination of these top features and the remaining 9 DES features (2 were already included in the top 40) was tested to evaluate whether the addition of more DES features improved accuracy. The resulting CA scores are shown in Table 15, with scores of at least 0.785 in bold. The final column contains the difference, with positive values indicating an improvement and negative values revealing a decrease in the models' accuracy.

In comparison to the first scenario (Section 4.1), four ML algorithms (DT, GB, LR, and NN) exhibited

Table 14: Scenario 2: CA scores for top 120 features in bundles of 10 across 10 ML models.

Classification Accuracy (CA) Scores												
Model	**10ft**	**20ft**	**30ft**	**40ft**	**50ft**	**60ft**	**70ft**	**80ft**	**90ft**	**100ft**	**110ft**	**120ft**
CN2	0.602	0.591	0.569	0.575	0.586	0.558	0.575	0.608	0.602	0.591	0.669	0.669
DT	0.630	0.652	0.663	0.652	0.707	0.691	0.663	0.652	0.641	0.619	0.635	0.635
AdaBoost	0.669	0.613	0.641	0.641	0.608	0.624	0.646	0.635	0.657	0.646	0.624	0.652
GB	0.724	*0.691*	0.685	0.691	0.685	0.663	0.685	0.674	0.702	0.713	0.702	*0.735*
NN	0.702	*0.729*	*0.740*	0.751	0.713	0.718	0.702	*0.729*	*0.729*	0.696	0.707	0.718
kNN	0.702	0.691	0.669	0.619	0.624	0.613	0.613	0.602	0.597	0.613	0.608	0.608
SVM	0.702	0.702	0.718	*0.740*	0.724	***0.735***	***0.740***	***0.740***	***0.746***	***0.740***	***0.729***	***0.740***
RF	*0.735*	0.707	0.702	0.713	***0.762***	0.691	0.718	0.724	0.691	0.724	0.713	*0.735*
LR	0.713	0.724	0.713	***0.785***	0.696	0.680	0.680	0.685	0.702	0.641	0.630	0.630
NB	***0.779***	***0.773***	***0.751***	0.718	*0.735*	0.718	0.674	0.674	0.657	0.669	0.669	0.663

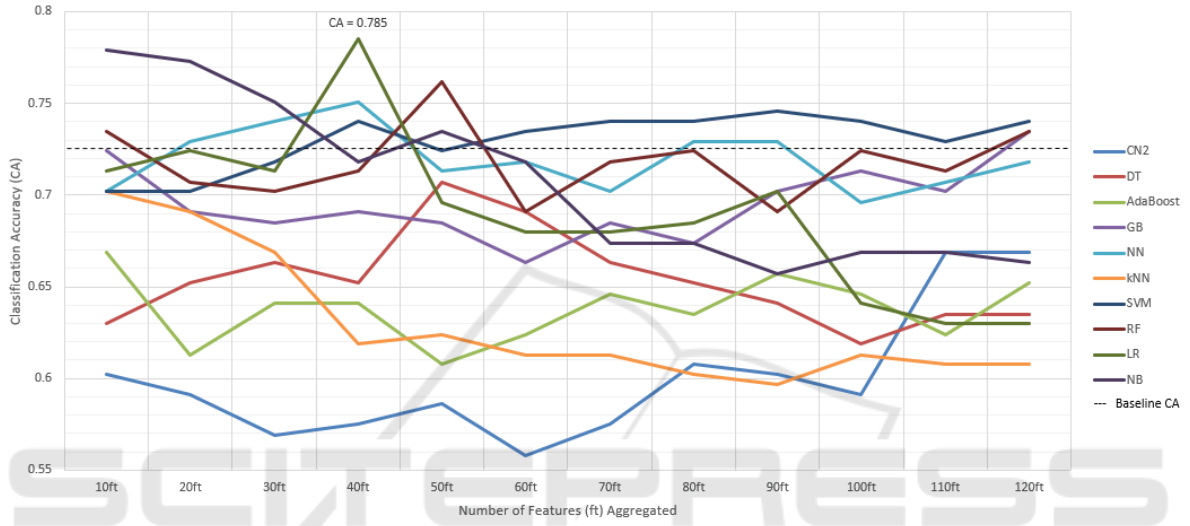


Figure 3: Scenario 2: Machine-learning algorithms performance based on feature aggregation with human rater classification.

Table 15: CA scores: top-ranked 40 plus remaining 9 DES features.

CA Model Performance Comparison			
Model	**w/40ft**	**w/40ft+9 DES**	**Difference**
DT	0.652	0.635	-1.7%
AdaBoost	0.641	0.663	+2.2%
CN2	0.575	0.591	+1.6%
GB	0.691	0.635	-5.6%
NB	**0.718**	**0.785**	**+6.7%**
RF	0.713	0.751	+3.8%
LR	0.785	0.713	-7.2%
kNN	0.619	0.669	+5%
SVM	0.740	0.746	+0.6%
NN	0.751	0.735	-1.6%

declines in CA scores, with LR experiencing the most significant drop of 7.2%. Conversely, models like AdaBoost, RF, kNN, SVM, and NN showed moderate improvements, averaging 2.64%, though their scores remained below the 0.785 threshold.

Notably, the NB model matched the highest CA score of 0.785 achieved by LR with the combined 40 features, consistent with its strong performance in the first scenario. While some models faced declines, including DES features continues to yield valuable insights into writing performance. This is further validated by the consistent performance of NB, a model recognized in the literature for its effectiveness in feature selection and pattern detection (Hirokawa, 2018).

5 CONCLUSIONS

This study aimed at refining English writing proficiency assessment and placement processes by leveraging a 300-essay corpus analyzed using a total of 445 linguistic features. These features were drawn predominantly from well-established platforms such as COH-METRIX and CTAP, but also included a set of 11 DES-specific features humanly devised and previously tested for this task, further emphasizing the importance of lexical and syntactic complexity analysis in placement accuracy (Stavans and Zadunaisky-Ehrlich, 2024). The 300 full-text samples were classified automatically by ACCUPLACER and, also, independently by two trained annotators into the following skill levels: Levels 1, 2, and College-level. This dual classification approach provided a comprehensive basis for comparing automated and human evaluations.

The study contributes to ongoing efforts in refining feature selection for text classification, supporting the argument that high-quality annotated corpora are essential for reliable automated assessments (Lee and Lee, 2023; Götz and Granger, 2024). Using ORANGE as a data mining tool, top-performing features were identified through Information Gain rankings and evaluated for their impact on classification accuracy (CA) across various machine learning (ML) models. This testing attempted to replicate the automatic classification process of ACCUPLACER, about which the literature currently provides limited information regarding its methodology and feature selection process.

DES features such as EXAMPLE and ORT ranked among the top 40 most discriminative features and offered granular insights into students' writing abilities in areas like argumentation and orthographic precision. They also improved classification accuracy in several ML models reinforcing prior work on the role of argumentation, lexical variation, and orthographic precision in proficiency assessment (Kosiewicz et al., 2023; Nygård and Hundal, 2024). Additionally, results align with research demonstrating the efficacy of feature selection in enhancing classification outcomes, as seen in studies utilizing COH-METRIX and CTAP for multilingual language assessment (Okinina et al., 2020; Leal et al., 2023; Akef et al., 2023).

The inclusion of human classifications in this study, validated through inter-rater reliability measures (Pearson and K-alpha inter-rater reliability coefficient), provided a critical benchmark for evaluating the efficacy of both the automated ACCUPLACER system and the machine learning models showcasing an alternative approach to reducing misclassification rates (Hughes and Li, 2019; Link and Koltovskaia, 2023). Unlike ACCUPLACER, which operates as a "black-box" system with limited transparency, human raters have the ability to capture nuanced linguistic features and transitions between proficiency levels. By combining these approaches, this study demonstrated improvements in CA, particularly with models like NB, which consistently matched or outperformed the baseline accuracy of 0.727 in both classification scenarios. These findings align with prior research on the effectiveness of ML in educational applications (Filighera et al., 2019).

The refined feature set and methodology presented in this study is a critical step in the broader process of advancing equity in educational outcomes by (i) improving student placement, (ii) reducing misclassification, and (iii) supporting targeted instructional interventions. Future efforts will continue to refine and validate these approaches, ensuring they align with the complex needs of DevEd in higher education.

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 profound gratitude to the dedicated annotators who participated in this task and the IT team whose expertise made the systematic analysis of the linguistic features presented in this paper possible. Their meticulous work and innovative approach have been instrumental in advancing our research.

REFERENCES

Akef, S., Mendes, A., Meurers, D., and Rebuschat, P. (2023). Linguistic complexity features for automatic Portuguese readability assessment. In *XXXIX Encontro Nacional da Associação Portuguesa de Linguística, Covilhã, Portugal, October 26–28, 2023, Proceedings 14*, pages 103–109. Associação Portuguesa de Linguística.

Bickerstaff, S., Beal, K., Raufman, J., Lewy, E. B., and Slaughter, A. (2022). Five principles for reforming Developmental Education: A review of the evidence. *Center for the Analysis of Postsecondary Readiness*, pages 1–8.

Chen, X. and Meurers, D. (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.

Cohen, J. (1988). *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Colucci Cante, L., D'Angelo, S., Di Martino, B., and Graziano, M. (2024). Text annotation tools: A comprehensive review and comparative analysis. In *International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 353–362. Springer.

Da Corte, M. and Baptista, J. (2022). A phraseology approach in developmental education placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86.

Da Corte, M. and Baptista, J. (2024a). Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti,

S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3046–3056, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024b). Enhancing writing proficiency classification in developmental education: The quest for accuracy. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024c). Leveraging NLP and machine learning for English (l1) writing assessment in developmental education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140.

Da Corte, M. and Baptista, J. (2025). Toward consistency in writing proficiency assessment: Mitigating classification variability in developmental education. In *Proceedings of CSEDU 2025*, Porto, Portugal. (to appear).

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

Duch, D., May, M., and George, S. (2024). Empowering students: A reflective learning analytics approach to enhance academic performance. In *16th International Conference on Computer Supported Education (CSEDU 2024)*, pages 385–396. SCITEPRESS-Science and Technology Publications.

Edgecombe, N. and Weiss, M. (2024). Promoting equity in Developmental Education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1.

Feller, D. P., Sabatini, J., and Magliano, J. P. (2024). Differentiating less-prepared from more-prepared college readers. *Discourse Processes*, pages 1–23.

Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 335–348. Springer.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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

Ganga, E. and Mazzariello, A. (2019). Modernzing college course placement by using multiple measures. *Education Commission of the States*, pages 1–9.

Giordano, J. B., Hassel, H., Heinert, J., and Phillips, C. (2024). *Reaching All Writers: A Pedagogical Guide for Evolving College Writing Classrooms*, chapter Chapter 2, pages 24–62. University Press of Colorado.

Götz, S. and Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1):1–38.

Hirokawa, S. (2018). Key attribute for predicting student academic performance. In *Proceedings of the 10th International Conference on Education Technology and Computers*, pages 308–313.

Huang, Z. (2023). An intelligent scoring system for English writing based on artificial intelligence and machine learning. *International Journal of System Assurance Engineering and Management*, pages 1–8.

Hughes, S. and Li, R. (2019). Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, 41:72–75.

Johnson, M. S., Liu, X., and McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3):338–361.

Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B., and Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and writing*, 30:1287–1310.

Kochmar, E., Gooding, S., and Shardlow, M. (2020). Detecting multiword expression type helps lexical complexity assessment. *arXiv preprint arXiv:2005.05692*.

Kosiewicz, H., Morales, C., and Cortes, K. E. (2023). The "missing English learner"' in higher education: How identification, assessment, and placement shape the educational outcomes of English learners in community colleges. In *Higher Education: Handbook of Theory and Research: Volume 39*, pages 1–55. Springer.

Kyle, K., Crossley, S. A., and Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 43(4):781–812.

Laporte, E. (2018). Choosing features for classifying multiword expressions. In Sailer, M. and Markantonatou, S., editors, *Multiword expressions: In-sights from a multi-lingual perspective*, pages 143–186. Language Science Press, Berlin.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–38.

Lee, B. W. and Lee, J. H.-J. (2023). Prompt-based learning for text readability assessment. In *In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19. Toronto, Canada. Association forComputational Linguistics.

Link, S. and Koltovskaia, S. (2023). *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham.

Ma, H., Wang, J., and He, L. (2023). Linguistic features distinguishing students' writing ability aligned with CEFR levels. *Applied Linguistics*, 45(4):637–657.

McNamara, D. S., Ozuru, Y., Graesser, A. C., and Louwerse, M. (2006). Validating CoH-Metrix. In *Proceedings of the 28th annual Conference of the Cognitive Science Society*, pages 573–578.

Nygård, M. and Hundal, A. K. (2024). Features of grammatical writing competence among early writers in a Norwegian school context. *Languages*, 9(1):29.

Okinina, N., Frey, J.-C., and Weiss, Z. (2020). CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020*, pages 7123–7131.

Pajila, P., Sheena, B., Gayathri, A., Aswini, J., Nalini, M., and R, S. S. (2023). A comprehensive survey on Naïve Bayes algorithm: Advantages, limitations and applications. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1228–1234.

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

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

Roscoe, R. D., Crossley, S. A., Snow, E. L., Varner, L. K., and McNamara, D. S. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. *The Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*, pages 393–398.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Stavans, A. and Zadunaisky-Ehrlich, S. (2024). Text structure as an indicator of the writing development of descriptive text quality. *Journal of Writing Research*, 15(3):463–496.

Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

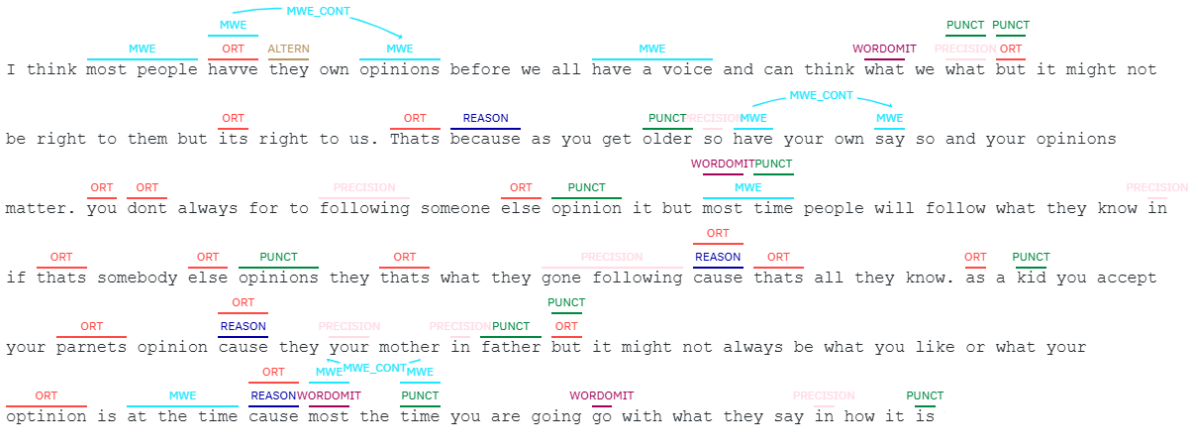
Vajjala, S. and Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Tetreault, J., Burstein, J., Kochmar, E., Leacock, C., and Yannakoudakis, H., editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Building Educational Applications using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

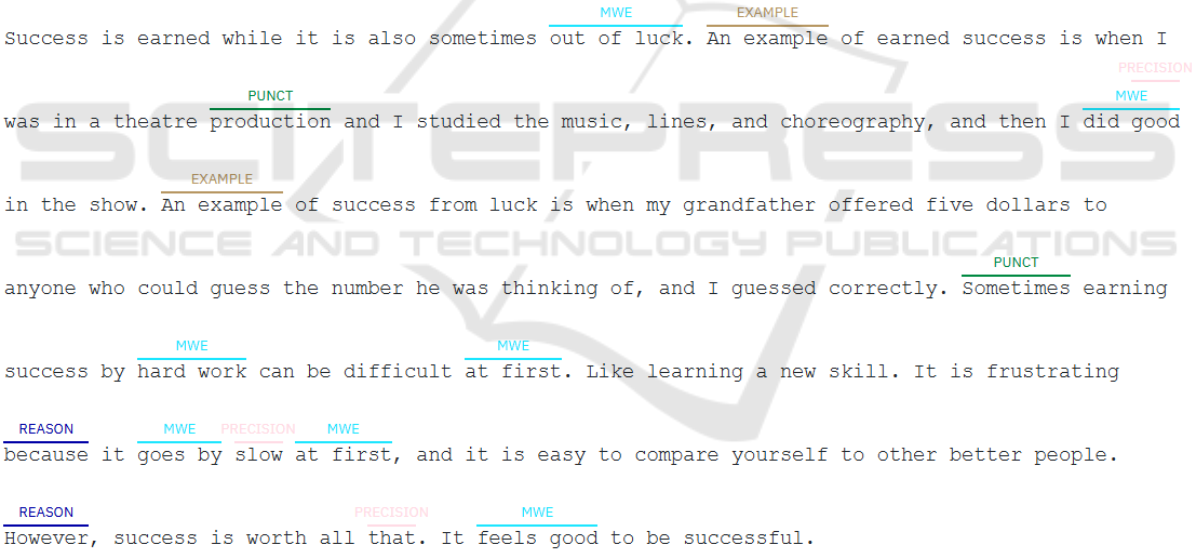
Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K. P., and François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

APPENDIX

Text ID: 66, Classified as Level 1 by both raters.



Text ID: 57, Classified as Level 2 by both raters.



Section 3

Beyond Accuracy: Supporting Consistency and Equity in Writing Assessment for DevEd Placement

Following the development of a linguistically grounded annotation framework in Section 1 and the validation of NLP-enhanced feature sets for improving ML models' classification accuracy in Section 2, *Section 3* now focuses on the consistency and equity of DevEd placement outcomes. Drawing from two studies, this section investigates whether ACCUPLACER's classifications consistently reflect students' actual writing performance and if it equitably classifies this performance across diverse student populations. It also addresses the third research question (RQ3) of this thesis: *What disparities in placement outcomes emerge across sociodemographic lines such as race and gender, and how can linguistically grounded features improve classification accuracy to promote more equitable DevEd placement*?

The first paper, titled *Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education* (Da Corte and Baptista (2025d)), revisits ACCUPLACER's reliability in comparison to human raters. While this particular study is closely related to (Da Corte and Baptista (2024b)), detailed and included in Section 2, its focus is different. This study addresses the following questions: (i) to what extent do ACCUPLACER's automated classifications align with expert human ratings in the assessment of developmental writing?, and (ii) how can human-informed descriptors enhance the accuracy and transparency of writing placement within DevEd?

Once again, the six textual descriptors used in the ACCUPLACER assessment (The College Board (2022)) were leveraged in the classification task devised in this study. The same original corpus of 100 text samples was used. All text samples were assigned to six pairs of raters, ensuring that at least two individuals reviewed each essay. The rating task was divided into two batches of 50 essays each, completed over the course of one month. This approach was adopted to ensure that raters had ample time to review the essays, prioritizing the quality of the evaluation process. The classification task consisted of two main steps:

(1) Assessing each essay according to six specific textual criteria, defined in the ACCUPLACER manual (The College Board, 2022, p. 24): (i) *mechanical conventions* (MC); (ii) *sentence variety & style* (SVS); (iii) *development & support* (DS); (iv) *organization & structure* (OS); (v) *purpose & focus* (PF); and (vi) *critical thinking* (CT). For each descriptor, raters used a simplified 4-point Likert scale[44]: 0 for *deficient*; 1 for *below average*; 2 for *above average*; and 3 for *outstanding*.

(2) Assigning an overall classification to each essay: DevEd Level 1, DevEd Level 2, College Level.

[44]This scale simplifies the more complex 8-point Likert scale currently used by ACCUPLACER (The College Board, 2022, p. 26).

In the first step (i), out of the 100 essays analyzed, 37 showed complete agreement between raters across all descriptors. Meanwhile, 63 essays exhibited discrepancies in at least one descriptor. The majority of these disagreements were relatively minor: 26 essays had a discrepancy in only one descriptor; 17 essays had discrepancies in two descriptors; 11 essays differed in three descriptors; and only nine essays showed disagreements across four descriptors. None received differing scores in five or six descriptors.

A closer examination of the descriptors that most frequently triggered disagreement revealed important patterns. The highest number of inconsistencies occurred in *development and support*, with 37 essays receiving differing scores. This result was followed by *sentence variety & style* (29 essays) and *purpose & focus* (19 essays). *Organization & structure* and *critical thinking* each had 16 cases of disagreement, while *mechanical conventions* showed the least variation, with only 12 essays rated differently.

Krippendorff's Alpha (K-alpha) inter-rater reliability coefficients were computed for the initial pair of raters using the ReCal-OIR[45] tool (Freelon (2013)), for ordinal data, to analyze the level of agreement for all six descriptors and the skill level. Four of the six descriptors, namely, *sentence variety & style* (α = 0.396), *purpose & focus* (α = 0.388), and *critical thinking* (α = 0.361), and *development & support* (α = 0.303), fell within the *fair* agreement range. In contrast, *mechanical conventions* (α = 0.566) and *organization & structure* (α = 0.433) achieved *moderate* agreement, suggesting that descriptor enhancements in these areas may have improved scoring consistency. As for the overall classification by level, interrater reliability reached a *moderate* agreement (α = 0.425), indicating moderate consistency in assigning skill levels, though still leaving room for refinement in operationalizing placement decisions.

Some of this reduced variability may be attributed to targeted refinements. As discussed in (Da Corte and Baptista (2024b)), the *mechanical conventions* descriptor was enhanced through the introduction of a numerical scale to quantify misspelled words. Similarly, the *organization & structure* descriptor was enhanced to better describe basic paragraph structure, with the following: *a text with an introduction and a thesis statement, supporting details, and a conclusion*. Although these ACCUPLACER descriptors were designed to reflect standard academic writing expectations, they seem not explicitly designed to capture in more detail the unique linguistic features of developing writers. This issue makes these descriptors inherently challenging to apply consistently, even for trained human raters. To resolve the rating mismatches for the 63 essays on which raters disagreed, a third independent rater, who had not previously evaluated the texts, was brought in to provide a supplementary assessment.

Enhancing linguistic descriptors is key to improving rater consistency and scoring precision, but equally important is understanding how these (enhanced) descriptors influence final skill-level classifications and, therefore, placement decisions. This leads to the completion of the second step (ii) in this classification task. When assigning the overall classification to each essay, based on the three levels, 68 essays received identical classification levels from both raters, while

[45] https://dfreelon.org/utils/recalfront/recal-oir/

32 essays were classified differently. The same third rater, who resolved the discrepancies with the linguistic descriptors, then evaluated these 32 essays to resolve the disagreements and establish the final placement.

ACCUPLACER's skill level classification used for the corpus sampling was balanced (50 essays per DevEd level); however, the human raters' classification resulted in 72 essays in DevEd Level 1 and 28 in DevEd Level 2. A confusion matrix comparing the two classification sources revealed a 22% misplacement rate by the automated system, primarily due to students' essays being classified as DevEd Level 2 by ACCUPLACER but rated as DevEd Level 1 by human annotators. This misplacement rate translates to approximately one in five students being placed in courses that are either too remedial or too advanced for their actual skill level. A Pearson score of ρ = 0.222 confirms the correlation between human raters and ACCUPLACER is *weak*.

The resulting, carefully vetted, and curated dataset of 100 essays produced using ACCUPLACER's descriptors and classified by skill level is presented in the paper and serves as a gold standard for writing placement in the DevEd context. This gold standard, which is included as a reference in **Appendix E**, was then used as ground truth for modeling the classification process using ML techniques and for understanding the role of each descriptor in refining the DevEd placement process. The ultimate goal behind leveraging ML techniques was threefold: (i) understand whether the system's classifications align with expert judgment; (ii) use the gold standard as a benchmark for modeling ACCUPLACER's DevEd classification task through ML algorithms, as its scoring algorithm remains undisclosed; and (iii) explore improvements for more transparent, systematic placement.

To better evaluate and refine placement performance, ML experiments using the ORANGE data-mining tool were conducted with a balanced set of 56 essays (28 per level) that contained the human-assigned ratings. Four frequently used ML models, namely, Random Forest (RF), Neural Network (NN), Decision Tree (DT), and Naïve Bayes (NB), were trained and tested on the six linguistic descriptors from ACCUPLACER, with the classification by level as the target variable. Considering the small size of the sample, the ORANGE's DATA SAMPLER widget was configured to partition it for 3-fold cross-validation, leaving two-thirds (37 essays) for training and one-third (19 essays) for testing purposes. Rather than testing a broader set of ML algorithms, this study focuses on the four models that previously demonstrated the highest accuracy in (Da Corte and Baptista (2024b)), thereby reinforcing continuity while also distinguishing this study's scope and emphasis from the earlier work.

The RF and NN models both achieved (*ex-aequo*) the highest classification accuracy (CA = 75.7%), with RF outperforming slightly in terms of Area Under the Curve metric (AUC = 85%)[46]. The other two models were DT and NB, with CA scores of 73% and 70.3%, respectively. Notably, the 75.7% CA achieved by RF aligns closely with the 72.7% baseline reported in (Da Corte and Baptista (2024c)). This similarity suggests that, although ACCUPLACER's descriptors may

[46]The AUC represents the likelihood that, when given a randomly selected positive and negative example, a ML algorithm assigns a higher score to the positive instance (Google for Developers (2025)).

be abstract and challenging for human raters to apply consistently, they demonstrate promise when refined with human input. With an improved classification rate approaching 80%, the enhanced descriptors now enable accurate placement for approximately eight out of ten students, representing an improvement over the previous rate of seven out of ten.

Feature ranking was then assessed with the Chi-square ranking method. Four descriptors emerged as most predictive: *development & support*, *organization & structure*, *mechanical conventions*, and *sentence variety & style*. A follow-up experiment was conducted using only these top-ranked features to reduce model complexity, minimize the risk of overfitting, and promote better generalization to unseen data. This refined configuration yielded improved model performance, with NB achieving an accuracy of 81.1%. This classification score represents an 11% gain over its initial performance, highlighting the benefits of feature selection and the algorithm's suitability for DevEd classification tasks, given its effectiveness, simplicity, and low computational cost (Pajila et al. (2023)).

The findings of this study suggest that placement accuracy can be improved by refining the use of linguistic descriptors, particularly through the integration of human-informed criteria into computational models, thereby contributing to fairer and more transparent placement outcomes.

While the first paper demonstrated the importance of refining descriptors to improve rater consistency and enhance classification accuracy, it primarily focused on the alignment between human and automated ratings. To build upon these findings, the second paper, titled *Beyond the Score: Exploring the Intersection Between Sociodemographics and Linguistic Features in English (L1) Writing Placement* (Da Corte and Baptista (2025a)), examines whether placement outcomes, produced by both ACCUPLACER and human raters, are related to or vary by sociodemographic factors such as gender and race. These demographic variables are deliberately excluded from ACCUPLACER placement decisions to mitigate bias and were also not accessible to annotators in this study. By introducing gender and race in the analysis, this second paper broadens the scope of this thesis, raising critical questions of fairness in how students are assessed and placed in developmental writing courses.

For this study, a corpus of 210 text samples was selected, along with a curated set of 40 linguistic features drawn from COH-METRIX, CTAP, and DevEd-specific sources. This dataset was used to explore how race and gender may intersect with placement decisions and influence classification outcomes. The refined sample of 210 reflects only those cases for which demographic data, specifically race and gender, were consistently and voluntarily reported to TCC. Although smaller in size, the resulting 210-text subsample remains representative of the 1,384 English (L1) students who took the ACCUPLACER test during the research period. Representativeness was again calculated using a simple index. By gender, females (1,384-sample: 62% $\rightarrow$ 210-subsample: 67%; RI = 0.919) and males (37% $\rightarrow$ 30%; RI = 0.811). By race, White students (41% $\rightarrow$ 42%; RI = 0.976), Black or African American students (15% $\rightarrow$ 17%; RI = 0.867), and American Indian students (10% $\rightarrow$ 11%; RI = 0.900).

As previously noted in (Da Corte and Baptista (2025c)), the original set of 300 essays was manually annotated by two trained raters who applied the respective DES features and independently assigned a skill-level classification. The set of 40 linguistic features used in this second study was also informed by the same work, which employed supervised ML techniques to evaluate the classification accuracy with a larger set of 445 features. Since not all features contributed equally to improving accuracy, the Information Gain scoring method was used here to identify the top-ranking features. Among the top 31[47] were nine features from COH-METRIX, 20 from CTAP, and two DES features, EXAMPLE and ORT, which ranked 9th and 31st, respectively. This ranking highlights the relevance of including human-devised descriptors in the classification task. To this end, the top 31 features were retained and paired with the remaining nine DES features, resulting in a total of 40 features, which are explicitly outlined in the paper. As supported by the findings in (Da Corte and Baptista (2025c)), incorporating the complete DES set enabled the Naïve Bayes (NB) model to achieve a classification accuracy of approximately 80%.

To examine whether placement outcomes, produced by both ACCUPLACER and human raters, are related to or vary by gender or race (or both), three commonly used statistical tests were run: (1) one-way ANOVA[48] to test whether the means of all 40 linguistic features differed significantly by gender (male, female, other/undisclosed), race (American Indian, Black or African American, White, and other/undisclosed), or both; (2) Tukey's Honestly Significant Difference (HSD)[49] test to identify pairwise group differences and measure effect sizes across gender and race categories; and (3) Chi-square test of independence to evaluate whether placement outcomes (ACCUPLACER vs. human-assigned) were associated with students' sociodemographic characteristics.

For these calculations, the conventional significance threshold of $\alpha = 0.05$ was applied. The one-way ANOVA was conducted to test for differences across 40 linguistic features among three gender groups (total $n = 210$). The results showed that for most features (36 out of 40), mean differences across gender groups were not statistically significant. The only *statistically significant* difference, as per the p thresholds in (Singh (2013)), was observed in the *positive connectives incidence* (p = 0.007), which refers to words that help build cohesion like *however* or *similarly*, by linking strings across texts. The remaining three features (ordered by p value) showed marginal effects, a result that warrants further investigation. These features were: *lexical richness hypergeometric distribution D (excluding punctuation and numbers)* (p = 0.051); *number of part of speech features: Existential there types* (p = 0.064); and *lexical richness: type token ratio (standardized type-token ratio for easy nouns from the New General Service List (NGSL)* (p = 0.094).

The one-way ANOVA was then conducted to test for differences across the same 40 linguistic features among four racial groups (total $n = 210$). The results showed that for most features

[47] A summary of the top 31 features is included in **Appendix F**, as a reference.

[48] Multi-way ANOVA was excluded due to sample size limitations in race × gender subgroups.

[49] Tukey's HSD is a post-hoc statistical method, used after an ANOVA shows significant differences between group means, to identify which groups differ from one another. (Abdi and Williams (2010))

(31 out of 40), mean differences across racial groups were not statistically significant. However, compared to the ANOVA gender-based results, a larger number of features showed either *significant* or *marginally significant* variation by race, with five of these being DES features. The three DES features showing *significant* variation are (ordered by p value): PUNCT (p = 0.000); WORDOMIT (p = 0.000); and PRECISION (p = 0.001). The next three features showing *statistically significant* variation are: *positive connectives incidence* (p = 0.005) *lexical sophistication feature: SUBTLEX logarithmic word frequency (average word type)* (p = 0.014); and *number of syntactic constituents: prenominal noun modifier* (p = 0.035). The last three features showed marginally significant variation: ORT (p = 0.057); REASON (p = 0.060); and *WordNet verb overlap* (p = 0.089). The fact that some DES features showed differences suggests that race-based variation may be more evident when using targeted, human-devised descriptors.

After the ANOVA calculations were completed, Tukey's HSD tests were conducted to determine whether any statistically significant pairwise differences existed between gender groups, focusing on the four features where ANOVA had revealed variation in group means. Tukey's HSD confirmed that, although ANOVA showed some variation in means, none of the gender-based pairwise comparisons reached significance. While female students showed higher use of *positive connectives*, the difference ($q_{\text{-Stat}} = 2.241 < q_{\text{critical}} = 2.788$) was not strong enough to be considered statistically reliable at this threshold. These findings should be interpreted with caution, given the limited sample size and unequal group distributions, which may have impacted the statistical power to detect more subtle effects.

Tukey's HSD tests were also conducted to determine whether any statistically significant pairwise differences existed between racial groups, focusing this time on the nine features where ANOVA had revealed variation in group means. For race, significant pairwise differences were noted with four linguistic features, three DES, namely, WORDOMIT, PUNCT, and PRECISION; and one from Coh-Metrix, *positive connectives incidence*. Black or African American students showed higher rates of omitted words (WORDOMIT) and less use of correct punctuation (PUNCT) in their written production. The imprecise use of a word, attending to its meaning (PRECISION) within a sentence or string of sentences, was also higher in Black or African American students than in the other two racial groups. On the contrary, Black or African American students showed a higher incidence of *positive Connectives* than American Indian students, suggesting greater use of linking expressions such as *furthermore* or *additionally*.

While some of these differences, particularly those involving the DES features, may be due to dialectal or regional variation, prior educational access, or even situational factors such as test anxiety during the Accuplacer exam, rather than actual deficits in academic ability (Pittman et al. (2024)), it is crucial to recognize two key points: (i) lexical and syntactic features remain essential indicators of college-level writing readiness; and (ii) academic standards must be upheld to ensure students are prepared to communicate effectively in their respective programs and beyond.

However, when such features are selected or weighted without attention to their broader sociolinguistic context, they may unintentionally disadvantage students from historically underrepresented backgrounds. While the ask is not to lower academic expectations, the results from this study are a call to (i) ensure that the features used to assess writing proficiency are both valid and fair, and to (ii) remain open to identifying additional features that are equally relevant, linguistically appropriate, and sensitive to diverse language experiences. In doing so, higher education institutions can move closer to relying on placement systems that uphold academic rigor while also promoting equity and inclusivity.

The final test run in this study used the Chi-square test to assess the association between gender and placement levels, both those automatically assigned by ACCUPLACER and those produced by human raters. For the automated classifications, the Chi-square value ($X^2 = 2.467$) calculated was below the critical Chi-square value ($X^2_{\text{critical}} = 5.991$) threshold, and the p value (0.291) exceeded the threshold of $\alpha = 0.05$. Similarly, for the human-assigned classifications, the Chi-square (X^2) value of 5.952 was also below the critical Chi-square value $X^2_{\text{critical}} = 9.488$, with a p-value of 0.203. In both cases $X^2 < X^2_{\text{critical}}$ and $p > \alpha$, indicating that the null hypothesis cannot be rejected. Therefore, based on these results, there is no *statistically significant* association between gender and the placement levels assigned by ACCUPLACER. The results obtained were very close to what would be expected by chance, suggesting gender did not influence placement outcomes in this sample.

When the Chi-square test was conducted to assess the association between race and placement levels, no *statistically significant* link between race and ACCUPLACER placement was found. Although the distribution of participants was not perfectly even, the differences were not strong enough to be considered meaningful. While this result might appear consistent with ACCUPLACER's stated goal of demographic neutrality, it does not preclude the possibility that the system's classifications still raise fairness concerns. A borderline result, which should be interpreted cautiously, was achieved with the human ratings. The Chi-square (X^2) value of 9.588 slightly exceeds the critical Chi-square value $X^2_{\text{critical}} = 9.488$, and the p-value = 0.048 is barely below the $\alpha = 0.05$. Although human raters, as previously addressed, did not have access to demographic data, the results of the Chi-square test indicate a possible association between race and human-assigned placements.

Upon analyzing each Chi-square contribution, the most significant discrepancies between observed and expected counts by race were examined as a final step to determine whether certain groups were over- or underrepresented at specific placement levels. Based on this sample, Black or African American students were overrepresented in DevEd Level 1 and underrepresented in College Level, while White students were more frequently placed above DevEd Level 1. This outcome may have been unintentionally influenced, in part, by the nature and definitions of certain DES features, raising questions about how rater judgments interact with linguistic patterns associated with race-based dialect differences. These placement patterns are consistent with higher education trends reported in the literature (Preston (2017); Morris-Compton (2021)),

which also highlight the low completion rates[50] of remedial courses among Black or African American students (Camardelle et al. (2022); Denison-Furness et al. (2022)).

The findings from this study reinforce the need not only to clarify placement guidelines and protocols, but also to strengthen the efforts to ensure pedagogical practices are supported by training aimed at reducing human biases. As with all studies included in this thesis, human expertise remains essential in assessing student writing, as human raters are better equipped to identify linguistic patterns that automated systems may overlook. These results align with the notion of employing a hybrid placement approach, incorporating human insights while leveraging the efficiency of NLP tools and ML techniques for diagnostic purposes.

References to the papers in order of presentation:

Paper 7:
Da Corte, M. and Baptista, J. (2025). **Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education**. In Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU, pages 139–150. INSTICC, SciTePress.

Paper 8:
Da Corte, M. and Baptista, J. (2025). **Beyond the Score: Exploring the Intersection Between Sociodemographics and Linguistic Features in English (L1) Writing Placement**. In Baptista, J. and Barateiro, J., editors, 14th Symposium on Languages, Applications and Technologies (SLATE 2025), volume 135 of Open Access Series in Informatics (OASIcs), pages 6:1–6:18, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[50]Percentage of students who complete a course with a grade of 'C' or higher, relative to all students enrolled initially, as defined by NCES.

Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education

Miguel Da Corte[1,2][a] and Jorge Baptista[1,2][b]
[1]*University of Algarve, Faro, Portugal*
[2]*INESC-ID Lisboa, Lisbon, Portugal*

Keywords: Developmental Education (DevEd), Automatic Writing Assessment Systems, English (L1) Writing Proficiency Assessment, Natural Language Processing (NLP), Machine-Learning (ML) Models.

Abstract: This study investigates the adequacy of Machine Learning (ML)-based systems, specifically ACCUPLACER, compared to human rater classifications within U.S. Developmental Education. A corpus of 100 essays was assessed by human raters using 6 linguistic descriptors, with each essay receiving a skill-level classification. These classifications were compared to those automatically generated by ACCUPLACER. Disagreements among raters were analyzed and resolved, producing a gold standard used as a benchmark for modeling ACCUPLACER'S classification task. A comparison of skill levels assigned by ACCUPLACER and humans revealed a "weak" Pearson correlation ($\rho = 0.22$), indicating a significant misplacement rate and raising important pedagogical and institutional concerns. Several ML algorithms were tested to replicate ACCUPLACER'S classification approach. Using the Chi-square (χ^2) method to rank the most predictive linguistic descriptors, Naïve Bayes achieved 81.1% accuracy with the top-four ranked features. These findings emphasize the importance of refining descriptors and incorporating human input into the training of automated ML systems. Additionally, the gold standard developed for the 6 linguistic descriptors and overall skill levels can be used to (i) assess and classify students' English (L1) writing proficiency more holistically and equitably; (ii) support future ML modeling tasks; and (iii) enhance both student outcomes and higher education efficiency.

1 INTRODUCTION AND OBJECTIVES

This study examines the adequacy of a machine-learning-based placement system, ACCUPLACER, and compares it to the classifications made by human raters within the context of U.S. Developmental Education (DevEd). By focusing on placement accuracy, this paper evaluates how reliable this system is and how effectively it supports higher education institutions in placing students into appropriate courses.

DevEd courses are designed to support students who are not yet college-ready, particularly by enhancing their English writing skills, ensuring they can gain access to and successfully participate in academic programs. Colleges typically assign students to either DevEd or college-level courses based on how they perform on standardized entrance exams. These decisions have significant consequences, as students with lower scores may need to complete one or two semesters of developmental, sometimes referred to as 'remedial,' coursework (Bickerstaff et al., 2022). Furthermore, the major concern among higher education institutions is the fact that current placement assessments are often unreliable in predicting students' performance, leading to incorrect placement for up to one-third of those who take the tests (Ganga and Mazzariello, 2019).

According to the National Center for Education Statistics (NCES)[1] approximately 3.6 million students graduated U.S. high school during the 2021-2022 academic year, with approximately 62%, ages 16 to 24, enrolling in colleges or universities as recently reported by the United States Bureau of Labor Statistics[2]. At Tulsa Community College[3], where this study was carried out, around 30% of incoming students are deemed not college-ready in more than

[a] https://orcid.org/0000-0001-8782-8377
[b] https://orcid.org/0000-0003-4603-4364

[1] https://nces.ed.gov
[2] https://www.bls.gov
[3] https://www.tulsacc.edu

Da Corte, M. and Baptista, J.
Toward Consistency in Writing Proficiency Assessment: Mitigating Classification Variability in Developmental Education.
DOI: 10.5220/0013353900003932
In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 139-150
ISBN: 978-989-758-746-7; ISSN: 2184-5026

one subject - English (reading and writing) and Math, highlighting the need for reliable placement mechanisms. Inaccurate placement in DevEd means students are assigned to courses beyond their writing abilities or to those that underestimate their skills, both of which have significant consequences for student outcomes and institutional effectiveness (Hughes and Li, 2019; Link and Koltovskaia, 2023; Edgecombe and Weiss, 2024).

This study aims to address these challenges by analyzing a corpus of 100 essays from college-intending students, evaluated according to six linguistic descriptors used by ACCUPLACER. The essays were classified by six human raters into one of three skill levels (DevEd Level 1, Level 2, or College Level) based on guidelines developed by two linguists using these descriptors. Subsequently, raters assessed the global level of the essays. This process unfolded in two phases over a period of one month.

A critical issue underlying this classification process lies in the nature of the descriptors themselves. Defined as high-order constructs, these descriptors encapsulate broad linguistic competencies but lack the granularity necessary to capture finer distinctions in writing proficiency. This study interrogates the extent to which these descriptors, particularly those associated with higher-order thinking, may require refinement to enhance assessment consistency and reliability. By examining how these constructs influence essay classification, the study seeks to provide insights into the alignment between linguistic features and skill-level assignments, ultimately informing the refinement of automated and human-scored placement decisions.

Given the background and context provided, this study has four objectives: (i) identify specific areas of disagreement in the application of six language-proficiency descriptors by human raters; (ii) compare the skill levels assigned by human raters with those automatically produced by ACCUPLACER; (iii) analyze and address discrepancies between proficiency descriptors and skill-level assignments; and (iv) develop a gold-standard corpus with a focus on achieving high inter-rater agreement.

Regarding its applicability to other languages, this study specifically focuses on English-language placement within U.S. community colleges. While extending the methodology to other languages is valuable, it falls outside this study's focus.

2 RELATED WORK

Current guidelines for DevEd assessment and placement vary widely across institutions and states, leading to inconsistencies in defining and categorizing proficient writing (Kopko et al., 2022). Automated systems like ACCUPLACER aim to reduce human biases and improve consistency in the assessment process (Link and Koltovskaia, 2023), but often lack the detailed insights—such as learners' early writing patterns—needed to shape effective instructional practices for developing writers (Da Corte and Baptista, 2024a).

Research calls for higher education institutions to turn to more comprehensive assessment tools to evaluate writing competencies, particularly for college readiness (Gallardo, 2021). As noted by (Lattek et al., 2024), "not every key competency can be assessed and measured using the same assessment method or instrument [...]; however, a suitable assignment or systematization is currently lacking," highlighting the need for refined scoring systems to ensure equitable and efficient access to higher education through linguistic development opportunities. Leveraging natural language processing (NLP) techniques (Link and Koltovskaia, 2023) could address these gaps and mitigate concerns about the accuracy and validity of current assessment methods (Barnett et al., 2020; Perelman, 2020).

Previous studies have aimed to detect textual patterns through computational methods, e.g., 'narrative style,' 'simple syntactic structures,' 'cohesive integration' (Dowell and Kovanovic, 2022), and to identify the linguistic features that are most predictive of accurate placement in Developmental Education (DevEd) (Da Corte and Baptista, 2024a). By focusing on descriptive features that better reflect native English proficiency, automatic classification systems have shown improvements in placement accuracy (Da Corte and Baptista, 2022), offering valuable insights into how to prepare students more effectively both linguistically and academically (Bickerstaff et al., 2022; Giordano et al., 2024).

Building on these findings, (Sghir et al., 2023; Duch et al., 2024) examined the use of ML algorithms to enhance the assessment of students' written productions and predict their performance and placement based on observable writing patterns (dos Santos and Junior, 2024). The overall findings indicate that well-known ML algorithms, such as Naïve Bayes, Neural Networks, and Random Forest, among others, are highly effective at pattern detection and feature selection (Hirokawa, 2018), and at predicting

students' outcomes, with classification accuracy rates above 90% (Duch et al., 2024).

As a result, refining proficiency descriptors, addressing discrepancies in their application by human raters and incorporating more descriptive linguistic features into the training of systems like ACCUPLACER could improve skill-level classification and better support DevEd students and faculty through learner corpora (Götz and Granger, 2024). Furthermore, developing a reliable, gold-standard corpus with high inter-rater agreement will establish a foundation for future assessments, supporting the overarching goals of this study.

3 METHODOLOGY

Through a systematic sampling method in accordance with the Institution's Review Board (IRB) protocols (ID #22-05[4]), this study ensured ethical and fair participant selection. It specifically focused on individuals who are educationally disadvantaged and adhered to guidelines that address the unique academic challenges faced by this group.

A rigorously selected corpus of 100 essays was assessed using six specific linguistic descriptors. Two linguists developed and tested classification guidelines based on these descriptors, and the essays were rated by trained human raters using a simplified 4-point Likert scale. Additionally, a pilot was conducted to validate the annotation approach, resulting in a comprehensive human-rated dataset benchmark. This golden standard dataset was then compared against the automated classification produced by ACCUPLACER.

3.1 Corpus

The current study builds on a previous classification task (Da Corte and Baptista, 2024b) that utilized a carefully selected corpus of 100 essays (27,916 tokens total)[5], written by college-intending students during the 2021-2023 academic years. Extracted from a larger pool of 290 essays within the standardized entrance exam database, these texts provided a robust foundation for assessing writing proficiency. Written in the Institution's proctored testing center without access to the Internet or editing tools, the essays were based on diverse prompts, such as the value of history, the acquisition of money, and the results of deception, designed to elicit analytical and reflective responses.

The selection process ensured that the corpus was representative of students' DevEd placement levels, categorized by ACCUPLACER as Level 1 or 2. The essays were balanced in level, as the classification by level was the target metric, and averaged 260 tokens each. Although small, this corpus serves as a critical foundation for analyzing the language proficiency of community college students. Demographic information (e.g., gender, race) was ignored at this stage.

This study's corpus specifically targets non-college level classifications to evaluate readiness for college and identify students needing DevEd support. Given the variability in DevEd guidelines across U.S. institutions, focusing on this segment is essential to refine placement accuracy through automated systems like ACCUPLACER, ensuring equitable access for academically underprepared students.

3.2 Classification Task Set-up

Before the classification task began, two linguists developed a set of guidelines[6], drawing on the six textual descriptors used in the ACCUPLACER assessment (The College Board, 2022). These guidelines were then tested in a pilot study, where a small selection of texts from the same exam database and timeframe was annotated. The focus was on identifying and categorizing relevant linguistic descriptors within DevEd.

In this classification task, all text samples (100) were randomly assigned to six pairs of raters, ensuring that each essay was reviewed by at least two individuals. To manage the substantial workload, the task was divided into two batches of 50 essays each, completed over the course of one month. Prior to this, the raters, who were familiar with DevEd and ACCUPLACER course placement in higher education institutions, received comprehensive training led by one of the authors of the annotation guidelines. This training covered the task's expectations, ethical considerations, an explanation of the guidelines, and the overall annotation procedures.

Following the training, the raters proceeded with the assessment, which involved two key steps:

(i) assessing each essay according to six specific textual criteria, defined in the ACCUPLACER manual (The College Board, 2022, p. 27):

(1) *Mechanical Conventions* (MC);
(2) *Sentence Variety and Style* (SVS);
(3) *Sentence Development and Support* (DS);
(4) *Organization and Structure* (OS);
(5) *Purpose and Focus* (PF); and
(6) *Critical Thinking* (CT).

[4]https://www.tulsacc.edu/irb

[5]Corpus dataset to be made available after paper publication.

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

For this evaluation, a simplified 4-point Likert scale[7] was applied to each criterion:

0 for *deficient*;
1 for *below average*;
2 for *above average*; and
3 for *outstanding*.

(ii) assigning an overall classification to each essay. Essays were classified based on definitions specifically adapted from the Institution's official course descriptions and curriculum:

DevEd Level 1 if essays demonstrated a need for improvement in general English usage, including grammar, spelling, punctuation, and the structure of sentences and paragraphs;

DevEd Level 2 if essays required targeted support in specific aspects of English, such as sentence structure, punctuation, editing, and revising; or

College level if essays were written accurately and displayed appropriate English usage at the college academic level.

4 EXPLORING VARIATIONS IN HUMAN RATER EVALUATIONS

4.1 Linguistic Descriptors

Each of the 100 essays received two scores, one from each rater, for the 6 linguistic descriptors analyzed. The assessments were completed on schedule, with raters self-reporting an average of 11 minutes per writing sample. Table 1 presents the number of essays that received differing scores in one or more descriptors.

Table 1: Essays that received differing scores in one or more descriptors.

Differing scored descriptors	Number of Essays
Essays with 0 differing scored descriptors	37
Essays with 1 differing scored descriptor	26
Essays with 2 differing scored descriptors	17
Essays with 3 differing scored descriptors	11
Essays with 4 differing scored descriptors	9
Essays with 5 or 6 differing scored descriptors	0
Total	100

Approximately 1/3 of the essays (37) received equal scores across all descriptors, indicating strong agreement between raters and suggesting a consistent application of the guidelines in these cases. In contrast, approximately 2/3 of the essays (63) had differing scores in one or more descriptors, hinting at some interpretative issues with the definitions provided.

Within a significant portion of this subset (63), 26 text samples had discrepancies in only 1 descriptor, while 17 of them exhibited differences in 2 descriptors. These essays likely represent cases where raters had minor disagreements, possibly due to subjective interpretation of specific descriptors. A total of 20 essays (combining those with 3 and 4 differing descriptors) revealed more pronounced disagreements among raters, indicating that scoring consistency decreases when assessing multiple linguistic aspects. No essays had discrepancies in 5 or 6 descriptors.

Upon close inspection of these 63 texts, three key themes are noted. Having a corpus of 100 essays and 6 linguistic descriptors results in a dataset with a total of 600 possible data points where differences between raters may occur. The focus here is on cases where Raters 1 and 2 provided contradictory scores, crossing the positive/negative boundary—where values of 0 or 1 signal a deficient or below-average text, and values of 2 or 3 signal an above-average or outstanding text.

Out of these 600 data points: 105 cases involved a one-point difference (e.g., 1 to 2) on the 4-point Likert scale; 23 cases involved a two-point difference (e.g., 0 to 2, 1 to 3); and 1 case, concerning the *Development and Support* descriptor, involved a three-point difference (3 to 0). In the remaining 471 instances, although some differing scores were observed within the same descriptors, they fell within either the negative (0, 1) or positive (2, 3) boundaries without crossing the +/- threshold.

To obtain a final score for each of the respective linguistic descriptors for the 63 essays included in Table 1, a third independent rater (Rater 3) was consulted. Rater 3 had not previously seen or evaluated the essays and provided an additional perspective to the assessment process. Results are included in Appendix 7.

4.1.1 Understanding Descriptor Discrepancies: Toward a Gold Standard

Although the descriptors used by ACCUPLACER align with general academic writing standards, they are not specifically tailored to the unique needs of DevEd students. These high-level descriptors are challenging to apply consistently, even for trained human raters, and based on current DevEd literature, lack grounding in detailed linguistic research addressing the specific requirements of DevEd contexts.

[7]This scale simplifies the more complex 8-point Likert scale currently used by ACCUPLACER (The College Board, 2022, p. 24).

The themes explored in Section 4.1 suggest that some descriptors, particularly those related to higher-order thinking, may be overly broad and could benefit from refinement. Narrowing their scope or providing more detailed guidelines could reduce discrepancies and improve assessment consistency. Table 2 focuses on analyzing which linguistic descriptors showed the greatest divergence in raters' scores, providing a basis for targeted improvements.

Table 2: Essays with differing scores per descriptor.

Descriptor	Number of Essays
Mechanical Conventions	12
Organization and Structure	16
Critical Thinking	16
Purpose and Focus	19
Sentence Variety and Style	29
Development and Support	37

During the pilot study mentioned in Section 3, refinements were made to the descriptors *Mechanical Conventions* and *Organization and Structure* to enhance their clarity and applicability. These efforts resulted in *Mechanical Conventions* achieving the fewest score discrepancies (12) among the raters. This improvement is attributed to specific enhancements made by two linguists to the descriptor's definition in the ACCUPLACER manual, including the introduction of a numerical scale for evaluating misspelled words:

0 (deficient) for 15 or more misspelled words;
1 (below average) for 8-14 misspelled words;
2 (above average) for 1-7 misspelled words; and
3 (outstanding) for no misspelled words.

However, unlike the refined scale for spelling errors, other grammatical features that could be incorporated into the *Mechanical Conventions* descriptor—such as word omission, word repetition, subject-verb disagreement (e.g., *we was* instead of *we were*), and punctuation misuse—lack similar metrics. These features will be considered in future refinements, as the ACCUPLACER manual provides no guidance on their inclusion in the assessment process.

Next are *Organization and Structure* and *Critical Thinking*, both with 16 essays each receiving differing scores. Minor enhancements were made to the *Organization and Structure* descriptor to include the fundamentals of paragraph composition: a clear introduction with a thesis statement, supporting statements, and a conclusion. This adjustment is believed to have enhanced the objectivity of the assessment. Texts relying on single lines or simple, non-multilayered paragraphs were rated as deficient, whereas those featuring well-structured paragraphs—with a clear introduction, at least two supporting statements, and a conclusion—were considered outstanding.

No enhancements were made to the *Critical Thinking* descriptor nor the remaining descriptors, as the intent was to closely assess the ACCUPLACER's classification criteria, and its reproducibility using human raters. *Critical Thinking*, despite being a complex and abstract feature, was measured based on constructs like *fairness*, *relevance*, *precision*, and *logic*, among others. Interestingly, it had the same number of essays scored differently as one of the descriptors whose definition was enhanced - *Organization and Structure*.

Under *Purpose and Focus*, 19 essays received differing scores. This feature evaluates how effectively a text presents information in a unified, coherent, and consistent manner. It also addresses the concept of *relevance*, thus partially overlapping with *Critical Thinking*. These aspects often require robust NLP tools for accurate detection and objective assessment.

The descriptors with the most scoring discrepancies were *Sentence Variety and Style* (29) and *Development and Support* (37). For *Sentence Variety and Style*, constructs such as *sentence length*, *vocabulary variety*, and *voice* are considered and can be easily quantified. For instance, sentence length exceeding 15 to 17 words—typically considered the standard for improved readability and comprehension (Matthews and Folivi, 2023)[8]—could serve as a basis for distinguishing between deficient and below-average texts and developing a numerical scale for this particular descriptor.

Regarding *Development and Support*, factors such as *point of view*, *coherent arguments*, and *evidence* are evaluated. While certain aspects, like the frequency of keywords introducing examples (e.g., *as proof*, *to give an idea*, *for example*) or reasoning (e.g., *because*, *although*, *consequently*), can be objectively measured, assessing a writer's point of view presents a significant challenge. This difficulty arises early in some students' academic journey, as articulating a coherent viewpoint often requires a level of maturity and experience gained through their studies.

While refining linguistic descriptors is essential for improving consistency and accuracy in rater assessments, it is equally important to consider how these descriptors contribute to the overall skill level classification. By examining how descriptors are applied to determine skill levels, discrepancies between human raters and ACCUPLACER classifications can be better understood, and that is the purpose of Section 4.2.

[8]https://readabilityguidelines.co.uk/

4.2 Skill Level

The overall skill level classification of the 100 essays is as follows: 68 essays received identical classification levels from both Rater 1 and Rater 2, while 32 essays were classified differently. For these 32 essays where discrepancies existed between human classifications, Rater 3, already mentioned in Section 4.1, was also tasked with independently providing a supplementary assessment to resolve the differences and confirm the skill level of the text samples. Still, for 2 essays, no agreement was reached, as each rater assigned a different skill level (Level 1, 2, and College Level). The final score was therefore determined by averaging the three ratings.

A custom scale based on the minimum and maximum average scores across all 6 linguistic descriptors was developed to further examine these differences. This scale provided a systematic framework for classifying texts into DevEd Level 1, DevEd Level 2, and College Level proficiency, with the following ranges:

Level 1 = [0.000 - 1.499];
Level 2 = [1.500 - 2.499]; and
College Level = [2.500 - 3.000].

Using this scale, the discrepancies noted in the assigned DevEd levels are summarized in Table 3.

Table 3: Discrepancies among assigned DevEd Levels.

Scale	Levels	Level 1	Level 2	College Level	Total
0.000 - 1.499	Level 1	61	**7**	0	68
1.500 - 2.499	Level 2	**11**	20	0	31
2.500 - 3	College Level	0	**1**	0	1
Total		72	28	0	100

From this summary, it was observed that: 7 texts with below-average and deficient scores (across the 6 linguistic descriptors) were assigned to DevEd Level 2; 11 texts with above-average and outstanding scores had been placed in DevEd Level 1; and 1 text with above-average and outstanding scores could have been placed in College-level writing but was deemed as a Level 2 text. The final human-assigned skill-level classifications are: 72 texts at Level 1 and 28 texts at Level 2.

Following this exploration of variability among human raters and the resulting assessment of all 100 text samples, a gold standard was established for non-college-level DevEd writing for the 6 linguistic descriptors and skill levels. Uniquely curated through human rater input, this gold standard is detailed in Appendix 7 and represents a significant advancement in creating a carefully vetted and reliable dataset for classification. It establishes a publicly available standard for this population that, to the best of the authors' knowledge, has not previously existed. Furthermore, this gold standard is essential for modeling the classification process using ML techniques and for understanding the role of each descriptor in refining the DevEd placement process, as detailed in Section 6.

5 INTER-RATER RELIABILITY AND QUALITY ASSURANCE

Based on the assessment results for the 6 linguistic descriptors and skill level classification, a rigorous quality assurance protocol was followed to evaluate inter-rater reliability, ensuring consistency and accuracy throughout the classification process.

Krippendorff's Alpha (K-alpha) inter-rater reliability coefficients were computed using the ReCal-OIR[9] tool (Freelon, 2013) for ordinal data. The aim was to analyze the level of agreement among the pair of raters for all 6 descriptors and the skill level. For the interpretation of K-alpha scores, the following agreement thresholds and interpretation guidelines, set forth by (Fleiss and Cohen, 1973), were followed:

below 0.20 - *slight* (Sl);
between 0.21 and 0.39 - *fair* (F);
between 0.40 and 0.59 - *moderate* (M);
between 0.60 and 0.79 - *substantial* (Sb);
above 0.80 - almost perfect (P);

Table 4 presents the results using both the granular 0-3 Likert scale, already presented in Section 3.2, and a binary scale, where 0 represents deficient or below-average scores, and 1 represents above-average or outstanding scores. The K-alpha interpretation thresholds are also provided. Items in bold indicate the top 3 linguistic descriptors with the highest reliability scores for each scale.

Table 4: K-Alpha scores for raters across 6 linguistic descriptors and skill levels: Likert vs. Binary scales.

Descriptors	0 - 3 scale	Interp.	0 - 1 scale	Interp.
Mechanical Conventions	**0.566**	M	**0.522**	M
Sentence Variety and Style	**0.396**	F	**0.352**	F
Development and Support	0.303	F	0.232	F
Organization and Structure	**0.433**	M	0.276	F
Purpose and Focus	0.388	F	0.213	F
Critical Thinking	0.361	F	**0.379**	F
Skill Level	0.425	M	0.413	M

K-alpha scores were anticipated to fall within the *slight* to *fair* agreement range, given the high-order nature and complexity of the linguistic descriptors as designed by ACCUPLACER. Results confirmed *fair* agreement for the descriptors *Sentence Variety and Style* (0.396), *Development and Support* (0.303), *Purpose and Focus* (0.388), and *Critical Thinking* (0.361)

[9]https://dfreelon.org/utils/recalfront/recal-oir/

on the 0–3 Likert scale. In contrast, *Mechanical Conventions* (0.566) and *Organization and Structure* (0.433) achieved *moderate* agreement.

On the binary scale, *Mechanical Conventions* was the only descriptor to achieve *moderate* agreement, representing the highest K-alpha scores overall across both scales. Skill level classification also demonstrated *moderate* agreement (0.425 on the Likert scale; 0.413 on the binary scale). The comparison between the binary and Likert scales revealed a "strong positive" Pearson correlation coefficient (Cohen, 1988) of $\rho = 0.754$.

The moderate scores for *Mechanical Conventions*, achieved, in part, due to the numerical scale for typos introduced, along with enhanced guidelines for *Organization and Structure* and self-developed skill level definitions aligned with the DevEd curriculum, suggest that further refining linguistic descriptors with specific, measurable criteria could lead to more objective and consistent evaluations, ultimately improving the placement process for DevEd students.

5.1 An Additional Quality Assurance Measure

Spearman rank correlation coefficients (ρ_s) were calculated based on the two raters' scores for each linguistic descriptor to assess their relative ranking across essays. Results are summarized in Table 5 and interpreted as follows (Mukaka, 2012):

0.90 to 1.00 *Very high positive correlation* (VHP);
0.70 to 0.90 *High positive correlation* (HP);
0.50 to 0.70 *Moderate positive correlation* (M);
0.30 to 0.50 *Low positive correlation* (LP);
0.00 to 0.30 *Negligible correlation* (N).

Table 5: Spearman rank correlation coefficients (ρ) for all 6 linguistic descriptors.

Rank	Linguistic Descriptor	Spearman Correlation	Interp.
1	Mechanical Conventions	0.632	M
2	Organization and Structure	0.521	M
3	Purpose and Focus	0.489	LP
4	Sentence Variety and Style	0.486	LP
5	Critical Thinking	0.471	LP
6	Development and Support	0.406	LP

Mechanical Conventions and *Organization and Structure* were the two descriptors with moderate positive correlation, suggesting a greater level of agreement between the two raters when evaluating these descriptors. The other four descriptors had similar low positive (ρ) scores, which can be attributed to their complexity and broadness, such as assessing the depth of ideas, the appropriateness of sentence structure, or the persuasiveness of an argument. These results seem to confirm the need for further refinements of these descriptors to improve inter-rater reliability.

6 MODELING ACCUPLACER'S CLASSIFICATION

This section investigates how the classification task performed by ACCUPLACER compares to that of human annotators when they apply the same purported linguistic criteria used as annotation guidelines. In alignment with the objectives of this study, outlined in Section 1, the experiments discussed here aim to evaluate the effectiveness of ACCUPLACER's criteria and assess whether human raters can consistently apply them. Eventually, the ultimate goal is to support a more systematic placement of students by enhancing such an automatic classification system.

It is important to note that ACCUPLACER's exact classification process is not publicly detailed, which presents a significant limitation. The system's classification operates as a "black box," and while a summary report is generated after the writing task assessment is completed, it does not detail scoring decisions, hindering educators and students from understanding why certain texts are classified as below college-level. This restricts opportunities for meaningful feedback and targeted improvements. To address these challenges and evaluate ACCUPLACER's performance, the gold standard scores established in Section 4 by human raters were used as a benchmark for modeling the system's DevEd classification task through Machine Learning (ML) experiments.

While ACCUPLACER's skill level classification used for the corpus sampling was balanced (50 essays per DevEd level), the human raters' classification results showed a different scenario (72 essays in DevEd Level 1 and 28 in DevEd Level 2). To further illustrate the discrepancies between ACCUPLACER and human classifications, a confusion matrix (Table 6) provides a detailed breakdown of specific instances where ACCUPLACER encounters challenges in accurately predicting proficiency levels.

Table 6: Accuplacer vs. Human Assessment: 100 Texts.

	Accuplacer L1	Accuplacer L2	Sum
Human L1	50	22	72
Human L2	0	28	28
Sum	50	50	100

Results from this matrix suggest a 22% misplacement rate by ACCUPLACER, where students were placed in DevEd Level 2 despite still needing significant development in their writing skills. The 22% error rate indicates that approximately 1 in 5 stu-

dents could be placed at an inappropriate level. This misclassification has serious pedagogical and institutional implications as (i) students placed below their skill level may face disengagement or frustration from redundant material, and (ii) students placed above their level might struggle, leading to lower success rates and higher attrition.

ACCUPLACER'S limitations and misalignments with human classifications were further validated through Pearson coefficient calculations, yielding a "weak" correlation of $\rho = 0.222$ and a comparable "slight" Krippendorff's Alpha inter-rater reliability coefficient of $k = 0.164$.

Before proceeding with the ML experiments outlined in Section 6.1, random resampling was conducted for Level 1 texts to ensure that the dataset was balanced according to the skill levels attributed by human raters. A total of 56 text samples (28 for each level—Levels 1 and 2) were then used.

6.1 Improving Placement Accuracy Through Machine Learning

The data mining tool ORANGE (Demšar et al., 2013)[10] was selected for analysis and modeling due to its comprehensive suite of popular and commonly used machine learning algorithms. The ORANGE toolkit website provides detailed documentation on classifier selection, definitions, and configurations, facilitating the evaluation of ACCUPLACER's performance relative to human classifications without the need to develop new models.

The workflow as shown in Figure 1 can be described as follows: the data is imported into Orange using the CSV File Import widget - one line per essay, 6 columns for the descriptors, plus a column with the overall skill level, used as the target variable. Data is then passed into the DATA SAMPLER widget, which, considering the small size of the sample, was configured to partition it for 3-fold cross-validation, leaving 2/3 (37 instances) for training and 1/3 (19 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, commonly used learning algorithms were chosen for their established performance in similar text classification tasks: Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Neural Network (NN). These algorithms were tested using the default hyperparameters provided by ORANGE. Classification Accuracy (CA) was the primary metric for analysis, while Area Under the Curve (AUC) was employed to differentiate between models with *ex-aequo* CA values.

While previous experiments conducted with the same corpus included six additional algorithms available in ORANGE—Adaptive Boosting, CN2 Rule Induction, Gradient Boosting, k-Nearest Neighbors, Logistic Regression, and Support Vector Machine—this study focused on the four models that delivered the most significant results in this context.

The results from this first part of the experiment are shown in Table 7.

Table 7: ML Algorithm CA w/ 6 linguistics descriptors.

Model	AUC	CA
Random Forest	**0.850**	**0.757**
Neural Network	0.741	**0.757**
Tree	0.722	0.730
Naïve Bayes	0.788	0.703

The order of the algorithms based on their CA scores is as follows: RF > NN > DT > NB. While RF and NN had *ex-aequo* CA scores, RF showed a higher AUC. The RF algorithm is frequently used in similar text classification tasks (Huang, 2023), thus this result is within expectations. The performance of the remainder algorithms is also not very far behind. A CA score of 0.757 suggests that, when using the 6 linguistic descriptors and the corresponding human-annotated scores replicating ACCUPLACER'S classification task, about 3 out of 10 students are inaccurately placed in DevEd courses. This CA value is comparable to the 0.727 CA baseline achieved by RF in a previous DevEd experiment (Da Corte and Baptista, 2024c). In that study, a large set of linguistic features were automatically extracted using Natural Language Processing (NLP) tools like CTAP[11] (Chen and Meurers, 2016), with approximately 300 focused on lexical patterns, such as lexical density and richness, and syntactic patterns, including syntactic complexity and referential cohesion, among others. Although this feature set is broader, the 0.727 CA baseline achieved with RF is only slightly lower than the 0.757 CA noted in this paper using just the six (slightly enhanced) descriptors from ACCUPLACER. This result highlights that, while ACCUPLACER'S constructs can be abstract and difficult for humans to apply in text assessment, they present an opportunity for increased accuracy when descriptors are enhanced with human input. The need for further refinement and improvement remains clear when compared to the more feature-rich approach in the earlier study.

[10] https://orangedatamining.com/

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

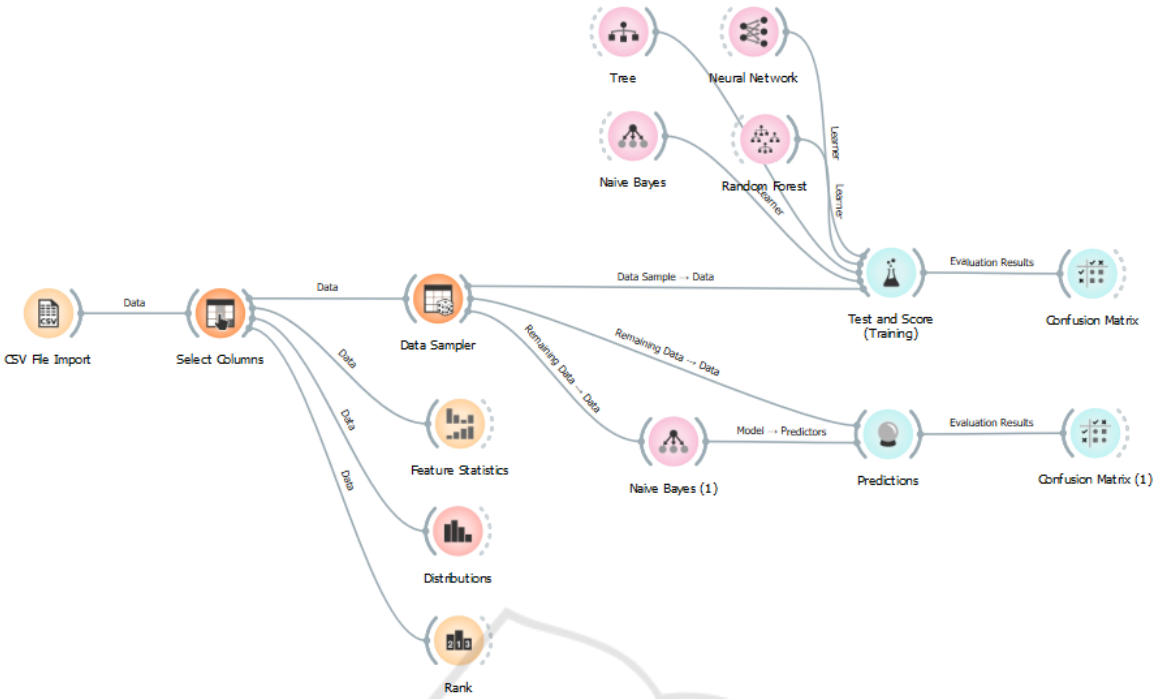


Figure 1: Orange Workflow Configuration for Model Training and Testing.

To examine how the ACCUPLACER'S descriptors contribute to the classification process, the built-in Information Gain and Chi-square (χ^2) ranking methods of Orange's Rank widget were compared to score those features. Figure 2 provides a visual representation of how the features ranked with both methods.

	#	Info. gain	χ^2
1	Development & Support	0.273	6.870
2	Organization & Structure	0.338	6.870
3	Mechanical Conventions	0.159	5.062
4	Sentence Variety & Style	0.218	4.263
5	Critical Thinking	0.187	4.050
6	Purpose & Focus	0.182	3.767

Figure 2: Descriptors ranked by Information Gain and Chi-square (χ^2) methods.

Development and Support and *Organization and Structure* ranked the highest with both ranking methods. As for *Mechanical Conventions* and *Sentence Variety and Style* are also the next highest ranking but in reverse order. *Critical Thinking* and *Purpose and Focus* ranked the lowest with both methods. The "very strong" Pearson correlation coefficient between the two scoring methods ($\rho = 0.824$) led to the adoption of the Chi-square (χ^2).

Using the ranking for feature selection, the top 4 descriptors, *Development and Support*, *Organization and Structure*, *Mechanical Conventions*, and *Sentence Variety and Style* were then used to classify the text samples again. This should prevent overfitting and enable the models to generalize better to unseen data. The results of this second experiment are summarized in Table 8.

Table 8: ML Algorithm CA with Top 4 Descriptors Ranked by Chi-square (χ^2).

Model	AUC	CA
Naïve Bayes	0.799	**0.811**
Neural Network	0.751	0.757
Tree	0.724	0.730
Random Forest	0.769	0.622

This time, the order of the algorithms based on their CA scores is as follows: NB > NN > DT > RF. The order is similar to the one from the first experiment, without feature selection, except that NB and RF swapped places, and now the NB achieved a higher AC score. Notably, NB demonstrated an improvement of nearly 11%. This can be viewed as an improvement in adequately placing 8 students out of 10 (instead of approximately 3 out of 10). This gain of accurately placing one more student, combined with NB's reputation for simplicity and minimal computational effort (Pajila et al., 2023), makes it a promising, well-suited algorithm for classification tasks, particularly in the context of DevEd.

7 CONCLUSIONS AND FUTURE WORK

This study evaluated the adequacy of machine learning (ML) based systems, particularly ACCUPLACER, in comparison to human rater classifications within a DevEd placement context. A corpus of 100 essays was assessed using ACCUPLACER'S 6 linguistic descriptors, with two—*Mechanical Conventions* and *Organization and Structure*—linguistically enhanced with quantifiable criteria, to improve inter-rater reliability. Human raters also classified the essays into two DevEd levels (Levels 1 and 2), and these classifications were compared to those assigned by ACCUPLACER.

Significant differences between human and automated classifications were noted, with ACCUPLACER presenting a 22% misplacement rate, compared to the human ratings. This result merits careful consideration. As with any human intelligence task (HIT), discrepancies in both the application of the 6 descriptors and the level assignments were carefully analyzed and resolved, leading to the production of a gold standard for classification—an important contribution that provides a curated dataset for future machine-learning modeling. This gold standard was subsequently used to mimic ACCUPLACER's task using ML algorithms, in which Random Forest achieved a classification accuracy (CA) of 0.757, comparable to a CA baseline of 0.727 obtained with a broader set of automatically extracted linguistic features and used in a prior study (Da Corte and Baptista, 2024c).

Despite ACCUPLACER'S constructs being abstract and challenging for human raters, the results demonstrate that high accuracy can still be achieved when these constructs are enhanced with human input. This was evident in the second experiment with the Naïve Bayes algorithm, which showed a performance improvement of nearly 11% (CA = 0.811). This translates to correctly placing approximately 8 out of 10 students in DevEd courses (versus approximately 3 out of 10). The ranking of descriptors using the Chi-square (χ^2) method further emphasized the importance of refining key features like *Development and Support* and *Organization and Structure*, which ranked highest.

Consequently, this study proposes to focus on refining (and only using) four key descriptors (instead of six) —*Development and Support*, *Organization and Structure*, *Mechanical Conventions*, and *Sentence Variety and Style*—by incorporating more precise linguistic features as criteria. The enhancements will include (i) *Development and Support*, incorporating keywords related to argumentation with reasons and examples to better evaluate how effectively a text presents and supports ideas; (ii) *Organization and Structure* more precisely described and measured with features like word omission, pronoun alternation, and enhanced paragraph composition criteria; (iii) *Mechanical Conventions*, which will identify orthographic features such as contractions, word boundary splits, and punctuation misuse; and (iv) *Sentence Variety and Style*, expanded to include grammatical and lexical-semantic features like word repetition, subject-verb agreement, word precision, and multiword expressions (MWE), which have been used to assess proficiency (Arnold et al., 2018).

These refinements will be tested on a larger corpus, with plans to incorporate 600 additional text samples (from the academic year 2023-2024) that will soon be available, including College Level data previously unavailable. At this stage, a two-step classification procedure is envisaged: (i) College/non-College, followed by (ii) DevEd Level-1/Level 2. Additionally, Large Language Models (LLMs), such as Generative Pre-trained Transformer (GPT), will be leveraged to see how feasible it will be to further align textual features with these refined descriptors and generate explanations for why certain essays meet or fail to meet specific criteria. This could potentially help resolve inter-rater conflicts and improve the overall classification process for DevEd placements, offering critical insights into both writing proficiency and curriculum design. Most importantly, these insights could be used to more effectively prepare and support students in their academic programs.

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 profound gratitude to the dedicated annotators who participated in this task and the IT team whose expertise made the systematic analysis of the linguistic features presented in this paper possible. Their meticulous work and innovative approach have been instrumental in advancing our research.

REFERENCES

Arnold, T., Ballier, N., Gaillat, T., and Lissón, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Barnett, E. A., Kopko, E., Cullinan, D., and Belfield, C. (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*.

Bickerstaff, S., Beal, K., Raufman, J., Lewy, E. B., and Slaughter, A. (2022). Five principles for reforming developmental education: A review of the evidence. *Center for the Analysis of Postsecondary Readiness*, page 1.

Chen, X. and Meurers, D. (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.

Cohen, J. (1988). *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Da Corte, M. and Baptista, J. (2022). A phraseology approach in developmental education placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86.

Da Corte, M. and Baptista, J. (2024a). Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3046–3056, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024b). Enhancing writing proficiency classification in developmental education: The quest for accuracy. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, Torino, Italia. ELRA and ICCL.

Da Corte, M. and Baptista, J. (2024c). Leveraging NLP and machine learning for English (l1) writing assessment in developmental education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

dos Santos, S. and Junior, G. (2024). Opportunities and challenges of AI to support student assessment in computing education: A systematic literature review. *CSEDU (2)*, pages 15–26.

Dowell, N. and Kovanovic, V. (2022). Modeling educational discourse with natural language processing. *Education*, 64:82.

Duch, D., May, M., and George, S. (2024). Empowering students: A reflective learning analytics approach to enhance academic performance. In *16th International Conference on Computer Supported Education (CSEDU 2024)*, pages 385–396. SCITEPRESS-Science and Technology Publications.

Edgecombe, N. and Weiss, M. (2024). Promoting equity in developmental education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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

Gallardo, K. (2021). *The Importance of Assessment Literacy: Formative and Summative Assessment Instruments and Techniques*, pages 3–25. Springer Singapore, Singapore.

Ganga, E. and Mazzariello, A. (2019). Modernzing college course placement by using multiple measures. *Education Commission of the States*, pages 1–9.

Giordano, J. B., Hassel, H., Heinert, J., and Phillips, C. (2024). *Reaching All Writers: A Pedagogical Guide for Evolving College Writing Classrooms*, chapter Chapter 2, pages 24–62. University Press of Colorado.

Götz, S. and Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1):1–38.

Hirokawa, S. (2018). Key attribute for predicting student academic performance. In *Proceedings of the 10th International Conference on Education Technology and Computers*, pages 308–313.

Huang, Z. (2023). An intelligent scoring system for English writing based on artificial intelligence and machine learning. *International Journal of System Assurance Engineering and Management*, pages 1–8.

Hughes, S. and Li, R. (2019). Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, 41:72–75.

Kopko, E., Brathwaite, J., and Raufman, J. (2022). The next phase of placement reform: Moving toward equity-centered practice. research brief. *Center for the Analysis of Postsecondary Readiness*.

Lattek, S. M., Rieckhoff, L., Völker, G., Langesee, L.-M., and Clauss, A. (2024). Systematization of competence assessment in higher education: Methods and instruments. In *CSEDU (2)*, pages 317–326.

Link, S. and Koltovskaia, S. (2023). *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham.

Matthews, N. and Folivi, F. (2023). Omit needless words: Sentence length perception. *PloS one*, 18(2):e0282146.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Pajila, P. B., Sheena, B. G., Gayathri, A., Aswini, J., Nalini, M., et al. (2023). A comprehensive survey on Naive Bayes algorithm: Advantages, limitations and applications. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1228–1234. IEEE.

Perelman, L. (2020). The BABEL generator and e-Rater: 21st Century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment*, 13(1).

Sghir, N., Adadi, A., and Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and information technologies*, 28(7):8299–8333.

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

APPENDIX

Gold standard for the 6 linguistic descriptors and skill levels with a corpus of 100 DevEd text samples.
Data presentation:
ID = document ID; **MC** = Mechanical Conventions; **SVS** = Sentence Variety & Style; **DS** = Development & Support; **OS** = Organization & Structure; **PF** = Purpose & Focus; **CT** = Critical Thinking; and **DevEd** = Development Education Level ("1" or "2").
Likert scale: **0** = deficient; **1** = below average; **2** = above average; and **3** = outstanding.

1,0,1,1,1,1,1,1
3,0,1,1,1,1,1,1
4,1,1,0,0,1,1,1
5,0,1,2,1,2,2,1
9,1,1,1,1,2,1,1
12,1,1,1,2,1,1,1
13,2,1,1,1,2,2,1
21,1,2,1,1,2,1,1
23A,0,0,1,0,1,0,1
23B,0,1,2,1,2,1,1
24A,2,2,2,2,2,1,2
24B,1,1,1,0,1,1,1
25,0,1,1,1,1,0,1
26,0,1,1,1,1,1,1
27,1,1,0,0,0,0,1
29A,0,1,1,0,1,0,1
29B,0,1,1,1,1,1,2
38,1,2,2,1,2,2,1
40A,0,1,2,1,2,2,1
40B,0,0,1,0,0,1,1
40C,1,1,1,1,2,2,1
45,2,2,0,1,1,1,1
48,0,1,1,0,1,1,1
49A,1,1,2,2,2,2,2
49B,1,1,1,0,1,1,1
51,1,2,2,1,2,1,1
55,1,1,1,1,1,1,1
56A,1,1,0,0,0,0,1
56B,1,1,1,1,1,1,1
59,2,1,1,1,1,2,1
60,1,1,1,0,1,1,1
61A,1,1,2,2,1,2,2
61B,0,1,1,1,1,1,2
69A,0,1,2,1,1,1,1
69B,1,1,1,1,2,1,1
70,1,0,0,1,1,1,1
73A,1,1,1,1,1,1,2
73B,2,1,1,1,2,1,2
73C,1,1,1,1,1,1,1
74,0,1,1,2,1,1,1
76,0,0,1,1,1,1,1
77,0,1,2,2,2,2,2
78,0,0,1,0,0,1,1
80A,1,2,2,2,2,1,2
80B,1,0,0,1,1,1,1
81,0,1,1,0,1,1,1
82,2,1,1,1,1,1,2
83,1,1,1,0,1,1,1
85,1,1,2,2,2,2,2
90A,1,2,2,2,2,2,2
90B,0,1,1,1,1,1,1
92A,0,1,0,1,0,0,1
92B,1,1,1,0,0,0,1
93A,0,1,2,2,1,2,2
93B,1,1,1,1,1,1,1
94,2,1,1,1,1,1,1
95A,1,1,1,2,2,1,1
95B,0,0,0,0,0,1,1
96,1,1,1,1,1,1,1
99,1,1,1,1,1,1,1
100,2,2,2,2,3,3,2
103,0,1,2,1,1,1,1
104A,1,1,1,1,1,2,1
104B,2,2,1,1,2,2,2
104C,1,2,1,1,1,1,2
106,1,1,2,2,2,2,1
110,1,2,1,1,1,1,1
113,0,2,1,0,1,1,1
115,0,1,1,0,1,1,1
116,1,1,1,1,1,1,1
118,1,1,1,1,1,1,1
119,2,3,2,2,3,2,2
120,2,1,0,1,2,2,1
123,3,2,1,1,1,1,1
124,1,2,1,1,2,1,1
125A,2,2,3,2,3,2,2
125B,1,1,1,1,1,2,1
125C,1,1,1,1,1,0,1
126,1,1,1,1,2,1,1
132,1,1,2,2,1,2,1
134,2,2,3,3,2,3,2
135,1,1,0,1,1,0,1
138,1,1,1,1,1,1,1
140,3,2,1,3,2,1,2
143A,2,2,3,2,2,2,2
143B,1,2,2,2,2,1,1
147,2,2,1,1,1,2,2
150,0,2,1,1,1,2,1
151A,1,2,1,1,2,1,1
151B,1,1,2,1,2,2,1
152,0,1,1,0,1,1,1
163,2,1,1,1,2,2,1
174,2,2,2,2,2,3,2
178,1,2,2,2,2,2,2
180,2,2,2,2,2,2,2
184,2,2,3,2,2,2,2
187A,1,2,2,2,2,2,1
193,2,3,2,2,3,2,2
198,2,1,1,2,2,1,2
199,3,2,2,1,3,1,1

Beyond the Score: Exploring the Intersection Between Sociodemographics and Linguistic Features in English (L1) Writing Placement

Miguel Da Corte
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Jorge Baptista
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Abstract

This study examines the intersection of sociodemographic characteristics, linguistic features, and writing placement outcomes at a community college in the United States of America. It focuses on 210 anonymized writing samples from native English speakers (L1) that were automatically classified by ACCUPLACER and independently assessed by two trained raters. Disparities across gender and race using 40 top-ranked linguistic features selected from Coh-Metrix, CTAP, and Developmental Education-Specific (DES) sets were analyzed. Three statistical tests were used: one-way ANOVA, Tukey's HSD, and Chi-square. ANOVA results showed racial differences in nine linguistic features, especially those tied to syntactic complexity, discourse markers, and lexical precision. Gender differences were more limited, with only one feature reaching significance (Positive Connectives, $p = 0.007$). Tukey's HSD pairwise tests showed no significant gender group variation but revealed sensitivity in DES features when comparing racial groups. Chi-square analysis indicated no significant association between gender and placement outcomes but suggested a possible link between race and human-assigned levels ($\chi^2 = 9.588$, $p = 0.048$). These findings suggest that while automated systems assess general writing skills, human-devised linguistic features and demographic insights can support more equitable placement practices for all students entering college-level programs.

2012 ACM Subject Classification Social and professional topics → Student assessment; Social and professional topics → Adult education; Computing methodologies → Language resources; Computing methodologies → Lexical semantics; Social and professional topics → Race and ethnicity; Social and professional topics → Men; Social and professional topics → Women

Keywords and phrases Developmental Education (DevEd), sociolinguistic variation, text classification, Machine Learning, placement equity

Digital Object Identifier 10.4230/OASIcs.SLATE.2025.6

Funding 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).

1 Introduction and Objectives

Community colleges play a vital role in expanding access to higher education opportunities for high school graduates across the United States of America [21]. According to this country's Department of Education[1] definitions, community colleges are institutions responsible for

[1] `https://www.ed.gov/higher-education` (Last accessed: 29th July 2025; all URLs in this paper were checked on this date.)

providing access to the first two years of a Bachelor's degree, certificates in specialized trades, and/or the academic preparation for occupational credentials for work-ready programs. Operating, in most cases, under an open-admission policy, these institutions are also equipped to provide the academic support needed for students who are linguistically underprepared to join and complete an academic program.

At Tulsa Community College (TCC)[2], where this study took place, students who require academic support in key areas, particularly writing, are placed in a foundational literacy program known as Developmental Education (DevEd). Depending on their demonstrated writing skills prior to beginning an academic program, students can be placed into one of the following three levels (adapted version of the institution's official course descriptions):

- **DevEd Level 1**: for support with frequent grammar, spelling, and punctuation issues, and lacking cohesion;
- **DevEd Level 2**: for support with some structural improvements still needing targeted support;
- **College Level**: no support needed; skills show academic-level writing with minimal errors.

DevEd placement is typically determined using automated text classification systems. At TCC, ACCUPLACER[3] [41] is the system currently used to assess students' writing and assign placement levels accordingly. Existing literature raises concerns about the accuracy and fairness of such systems. For example, in [16] it was mentioned that automated placement tools misplace approximately one-third of students, either assigning them to courses that exceed their preparedness (overplacement) or to levels that do not reflect their actual skills (underplacement), raising both pedagogical and ethical concerns.

Placement decisions often show a mismatch between individual student data and overall institutional demographics, posing questions about whether those placed into non-college-level programs (e.g., DevEd) truly reflect the broader student population at their institutions [19, 35]. A contributing factor to these concerns is that, although some sociodemographic information is voluntarily self-disclosed by test-takers, most placement algorithms are designed to operate without reference to these factors (e.g., native language, race, gender) in an effort to ensure neutrality and avoid bias. However, excluding this information may limit the understanding of the underlying sociolinguistic context that shapes students' writing development; thus, potentially contributing to disparities in placement outcomes across different student groups [2, 31].

To further explore this gap, we investigate the relationship between sociodemographic variables, linguistic features in student writing, and placement outcomes. Our analysis draws on a corpus of 210 anonymized essays, randomly selected from 1,384 participants who self-identified as native English (L1) speakers, with each essay automatically and manually assessed for skill level. To validate linguistic differences and placement patterns, we employ a suite of statistical techniques, including one-way ANOVA, Tukey's Honestly Significant Difference (HSD) test [1, 36], and Chi-square tests of independence.

While a range of demographic characteristics exists (e.g., age, first-generation status, academic load), this study focuses specifically on gender and race for two reasons. First, these variables were consistently and reliably self-reported across both institutional and

[2] https://www.tulsacc.edu/
[3] https://accuplacer.collegeboard.org/

testing datasets, ensuring compatibility and completeness for statistical analysis. Second, other variables – such as age or socioeconomic status – were either unavailable, inconsistently reported, or insufficiently balanced to support valid statistical comparisons.

Grounded within the language resources and corpus linguistics domain, this study aims to contribute to the development of support tools for language teaching and assessment. It also connects with Machine Learning (ML) applications in Natural Language Processing and addresses ethical dimensions surrounding the fairness and transparency of automated placement systems. To support this investigation, we propose the following research questions:

1. How do the sociodemographic characteristics of the students selected for analysis align with the broader population of ACCUPLACER test-takers?
2. To what extent do gender and race correlate with the use of linguistic features in student writing at the developmental and college-entry levels?
3. Are there significant associations between these demographic variables (gender and race) and the placement outcomes – both automated (ACCUPLACER) and human-assigned?

Our paper is organized as follows: Section 2 reviews existing research on writing assessment and placement, linguistic features, and demographic disparities. Section 3 describes the institutional context, participant demographics, corpus, feature selection, and statistical procedures. Section 4 presents findings from the ANOVA and Tukey's HSD analyses. Section 5 examines associations between demographic variables and placement outcomes, both through ACCUPLACER and human assessment. We conclude by summarizing our key findings.

2 Related Work

Students' writing skills and placement into college-level or remedial courses are often assessed through standardized systems like ACCUPLACER. Existing literature raises concerns about the validity and fairness of such instruments [5, 14, 26], particularly given the disproportionate placement of students from historically marginalized groups into non-college level or DevEd courses [3, 35]. Placement into DevEd courses has been consistently associated with lower rates of retention[4] and completion[6], posing questions about the economical and social consequences of automated placement decisions [25, 30].

While systems like ACCUPLACER aim to reduce human biases by standardizing the assessment process, they often struggle to capture subtle indicators of students' early writing development. These "black box" tools typically emphasize surface-level correctness over deeper communicative competence, limiting their ability to assess whether students can meet academic writing demands [31]. As a result, students often leave the placement process without a clear understanding of what constitutes college-readiness in writing [18] or what traits of their written production may require remediation [29]. Researchers have argued that placement practices must move beyond raw scores to account for their broader consequences [22], as achievement gaps continue to persist across demographic groups. To help uncover potential mechanisms behind disproportionate placement, we explore the role of certain linguistic features, particularly those associated with developing writers in DevEd contexts, by examining how these features differ across racial groups and are also associated with placement levels.

[4] Percentage of first-time, full-time students who return the following year to continue with their academic program, according to the National Center for Education Statistics (NCES).[5]

[5] `https://nces.ed.gov/`

[6] Percentage of students who successfully complete a course with a grade of "C" or higher, relative to all students originally enrolled, as defined by NCES.

Recent studies further highlight disparities in writing placement outcomes by sex and racial background. For example, [6] reports that female students often outperform male peers in writing, while students from racially and minority groups, particularly Black students, tend to score lower on standardized assessments and, once placed in developmental courses, their completion rate is significantly low [13, 30]. Persistently low success rates, especially among Black students, suggest that DevEd courses may not adequately address students' linguistic needs. Equitable placement must therefore be paired with instruction that is both culturally responsive and linguistically informed. To address these challenges, assessment frameworks should integrate demographic considerations to: (i) prioritize direct measures of writing abilities to better understand students' linguistic diversity [27]; (ii) eliminate systemic barriers in classification, placement, and pedagogy that impede student success [25, 40]; and (iii) provide better (and more accurate) support for linguistically underprepared students [33, 39].

Efforts to improve transparency in writing placement tasks through linguistic feature analysis have gained momentum in recent years [24, 31]. Prior studies [10, 12] show that high-quality annotated corpora and careful feature selection significantly enhance the reliability of automated classification systems [17]. Human annotations, in particular, provide fine-grained insights into students' writing strengths and weaknesses, yielding improved classification outcomes across multiple ML models [21, 23, 32].

These findings align with broader research advocating for the integration of feature-rich platforms [42], such as Coh-Metrix[7] [28] and CTAP[8] [7], to enhance writing assessment practices [37]. Refining linguistic feature sets represents a critical step toward advancing placement equity [15], reducing misclassification risks, and supporting more targeted instructional interventions for students, including those enrolled in DevEd programs. This study contributes to these goals by offering new insights into placement outcomes through a linguistically informed lens.

3 Methodology

3.1 Institutional Demographics

This study builds on prior research into native English (L1) writing proficiency and placement practices [11], focusing on pre-enrollment students from the 2021–2022 and 2023–2024 academic years. In 2023–2024, TCC reported an unduplicated headcount of 14,538 students. As shown in Figure 1, White students made up nearly half the population, followed by Hispanic/Latino, Black or African American, and American Indian or Alaska Native students.

In terms of gender and academic load, as illustrated in Figure 2, approximately two-thirds of students were female and one-third male. Regarding academic load, 69% were enrolled part-time (taking about two courses per semester), while 31% were enrolled full-time (taking four to five courses). The average student age was 23, with 57% aged 24 or younger. Additionally, about 24% was first-generation college students, defined by NCES as those whose parents did not complete a college degree.

Given this demographic profile, writing placement and literacy emerge as central concerns, with nearly 50% of students placed into at least one DevEd course. The average pass rate for DevEd Levels 1 and 2 in 2023–2024 was 56%. Meanwhile, TCC's retention rate is approximately 65%, and the graduation rate for full-time students stands at 27%. These metrics are critical as institutions work to support student completion and transition into the workforce.

[7] https://soletlab.asu.edu/coh-metrix/
[8] https://sifnos.iwm-tuebingen.de/ctap/

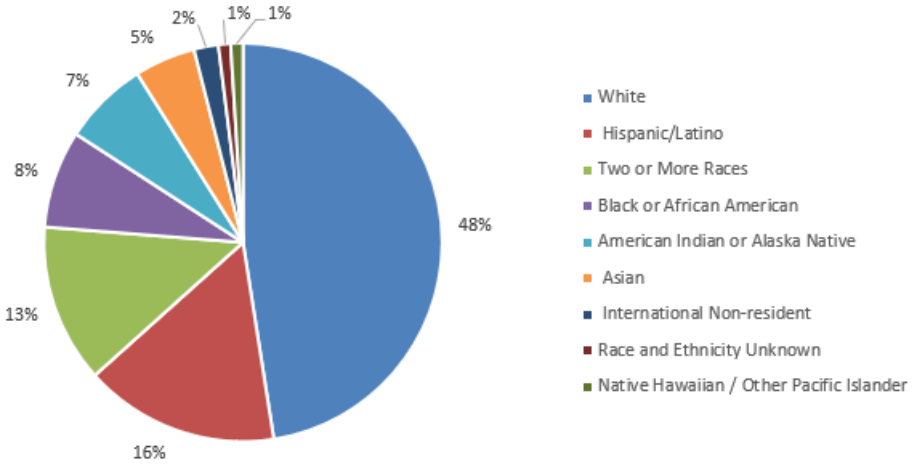


Figure 1 Distribution of the overall student population by race.

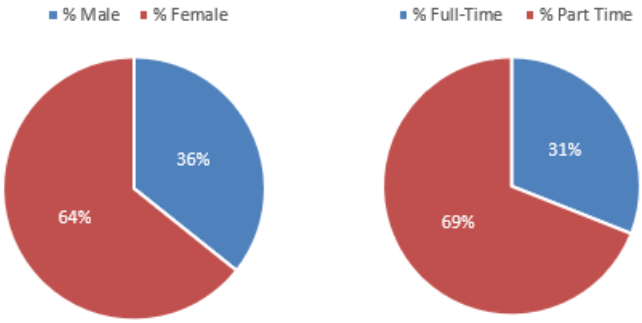


Figure 2 Distribution of the overall student population by gender and academic load.

3.2 Participants Demographics and Corpus Selection

To evaluate the representativeness of our sample relative to the broader ACCUPLACER test-taking population, as guided by our first research question, we examined institutional data from students who completed the ACCUPLACER exam during the same time period. Table 1 summarizes this data by gender, race, and native language.

A total of 2,003 students (unduplicated count) completed the ACCUPLACER test, producing a written sample that was automatically assessed and classified by proficiency level. In terms of gender, the test-taking population was composed of 62% females and 37% males, closely mirroring the broader institutional demographics. Regarding race, approximately 41% of students self-identified as White, followed by Black or African American (15%), American Indian (10%), and Latino/Hispanic (9%) students, proportions that closely align with the overall student population. First-language data further indicated that 69% of test-takers reported English as their native language, comparable to the 66% institutional ratio.

Building on prior investigations[9], this study narrows its analysis to the 1,384 participants who self-identified as native English (L1) speakers. Placement-level breakdowns for this group were available by gender only and are summarized in Table 2.

[9] An earlier phase of this research analyzed a random subset of 450 essays (from the 2,003 total); demographic data was not examined as placement was the focus.

Table 1 Demographic overview of ACCUPLACER test-takers during the 2021-2022 and 2023-2024 academic years.

Demographics	Group	Test-Takers	%
Gender	Female	1,244	62%
	Male	733	37%
	Other/Undisclosed	26	1%
	Total	**2,003**	**100%**
Race	American Indian	208	10%
	Black or African American	301	15%
	White	813	41%
	Latino/Hispanic	180	9%
	Other/Undisclosed	501	25%
	Total	**2,003**	100%
Native Language	English (L1)	**1,384**	**69%**
	Non-English	619	31%
	Total	**2,003**	**100%**

Table 2 ACCUPLACER placement by gender of English L1 test-takers.

Placement	Male	%	Female	%	Other/Undisclosed	%	Total
DevEd Level 1	24	21.82%	39	35.45%	2	1.82%	65
DevEd Level 2	225	22.59%	443	44.48%	5	0.50%	673
College Level	252	28.25%	378	42.38%	12	1.35%	642
Score not available	1	20.00%	3	60.00%	0	0.00%	4
Total	502	25.06%	863	43.09%	19	0.95%	1,384

From this table, we observed that approximately 53% of native English speakers were placed into one of the two DevEd levels (Level 1 or Level 2), with a higher concentration in Level 2. Placement into DevEd was slightly more common among female students. Notably, this 53% placement rate closely mirrors the broader institutional trend of approximately 50% outlined in Section 3.1.

To conduct a more detailed analysis, a subsample of 300 essays was randomly selected from the 1,384 native English speakers (≈22%). Within this subsample, complete and consistent demographic information was available for 210 participants, which formed the final dataset for this investigation. Among these participants, three racial groups are represented: American Indian (33; ≈16%), Black or African American (51; ≈24%), and White (125; ≈60%). One additional participant did not disclose their race but self-reported English as their native language (L1). Overall, the racial distribution in this subsample closely reflects both institutional demographics and those of the broader ACCUPLACER test-taking population.

Using the same placement classification criteria, we categorized these 210 participants into developmental and college-level groups, as summarized in Table 3.

Table 3 ACCUPLACER placement by gender; final dataset (n=210).

Placement	Male	%	Female	%	Other/Undisclosed	%	Total
DevEd Level 1	21	21%	33	33%	1	1%	**55**
DevEd Level 2	20	20%	57	57%	2	2%	**79**
College Level	23	23%	51	51%	2	2%	*76*
Total	64	21%	141	47%	5	2%	210

From this table, it is observed that students classified as DevEd (in bold) accounted for approximately 64% of the placements within the subsample (210 participants), while

those deemed proficient in writing (College Level; in italics) made up approximately 36%. Although smaller in size, this dataset, anchored in consistent demographic reporting, provides a foundation for the analyses examining the intersections between sociodemographic variables and linguistic features.

It is pertinent to note that all writing samples analyzed in this study were drawn from the institution's official entrance exam database. The writing task required students to compose an argumentative response (in English) to a standardized prompt encouraging reflection on a social, educational, or ethical issue. Students were instructed to complete the task in a proctored environment, in a single sitting, and without access to dictionaries, the internet, or AI tools.

3.3 Feature Selection

In addition to the ACCUPLACER classifications, all 210 essays with complete demographic information were manually annotated and evaluated for skill level by two trained raters following classification guidelines developed by the authors [8, 9]. To mitigate potential bias, multiple safeguards were implemented: all essays were anonymized, raters were blinded to students' demographic information, and a calibration session was conducted prior to annotation to ensure scoring consistency. In alignment with TCC's Review Board protocols, raters had no access to race, gender, or other sociodemographic data at any stage. While these measures reduced the risk of bias, we acknowledge that implicit bias may still persist.

The annotation process involved marking each text with a set of 11 Developmental Education-Specific (DES) linguistic features. These features captured both negative patterns (errors and deviations from proficiency) and positive patterns (indicators of advanced language use). They were organized into four clusters – Orthographic (ORT), Grammatical (GRAMM), Lexical and Semantic (LEXSEM), and Discursive (DISC), each reflecting a critical dimension of foundational writing proficiency, as summarized in Table 4.

Table 4 DevEd-specific (DES) features summary.

Pattern Description	Features Clustered
Orthographic patterns (ORT): representing the foundational language skills needed to represent words and phrases.	(-) Grapheme (addition, omission, transposition, and capitalization) (ORT) (-) Word split (WORDSPLIT) (-) Punctuation used & Contractions (PUNCT)
Grammatical patterns (GRAMM): evidencing the quality of text production.	(-) Word omitted (WORDOMIT) (-) Word repetition (WORDREPT) (-) Verb agreement (VAGREE) (-) Pronoun-alternation referential (ALTERN)
Lexical & Semantic patterns (LEXSEM): contributing to the structuring of a writer's discourse.	(-) Word precision (PRECISION) (+) Multiword expressions (MWE)
Discursive patterns (DISC): exhibiting the writer's ability to produce extended discourse.	(+) Argumentation with reason (REASON) (+) Argumentation with example (EXAMPLE)

Beyond the DES features, additional linguistic features were integrated from prior work [11], which included 106 from COH-METRIX and 328 from CTAP. Although 434 features[10] were originally extracted, only a subset was used in this study. Using Information Gain rankings from ML experiments, the most predictive features of placement were selected. Notably, two DES features – *EXAMPLE* and *ORT* – ranked 9th and 31st, respectively.

To emphasize the value of human-derived features, the remaining 9 DES features were retained, resulting in a final feature set of 40, as observed in Table 5. These DES features

[10] A detailed description of these features can be found in the documentation of these tools.

Table 5 Summary of highest-ranked features by Information Gain scores.

Rank	Source	Features	Info. gain
1	COH-Metrix	Paragraph length, number of sentences in a paragraph, mean	0.144
2	COH-Metrix	Lexical diversity, VOCD, all words	0.142
3	CTAP	Lexical Sophistication: Easy noun tokens (NGSL)	0.124
4	COH-Metrix	LSA given/new, sentences, mean	0.124
5	CTAP	Lexical Richness: Type Token Ratio (STTR NGSLeasy Nouns)	0.118
6	COH-Metrix	Lexical diversity, MTLD, all words	0.115
7	CTAP	Number of POS Feature: Plural noun Types	0.109
8	COH-Metrix	LSA given/new, sentences, standard deviation	0.109
9	**DES**	**EXAMPLE**	**0.103**
10	CTAP	POS Density Feature: Existential There	0.102
11	COH-Metrix	Positive connectives incidence	0.100
12	COH-Metrix	LSA overlap, adjacent sentences, mean	0.100
13	CTAP	Number of POS Feature: Existential there Types	0.099
14	CTAP	Number of POS Feature: Preposition Types	0.099
15	COH-Metrix	WordNet verb overlap	0.098
16	CTAP	Number of Syntactic Constituents: Postnominal Noun Modifier	0.097
17	CTAP	Number of Word Types (including Punctuation and Numbers)	0.097
18	CTAP	Lexical Sophistication Feature: SUBTLEX Logarithmic Word Frequency (AW Type)	0.096
19	CTAP	Number of POS Feature: Existential there Tokens	0.096
20	CTAP	Lexical Sophistication: Easy noun types (NGSL)	0.094
21	CTAP	Number of POS Feature: Verbs in past participle form Types	0.092
22	COH-Metrix	LSA verb overlap	0.092
23	CTAP	Number of Unique Words	0.092
24	CTAP	Number of Tokens with More Than 2 Syllables	0.090
25	CTAP	Number of Word Types with More Than 2 Syllables	0.086
26	CTAP	Lexical Richness: HDD (excluding punctuation and numbers)	0.086
27	CTAP	POS Density Feature: Possessive Ending	0.086
28	CTAP	Number of Syntactic Constituents: Complex Noun Phrase	0.084
29	CTAP	Number of POS Feature: Plural noun Tokens	0.083
30	CTAP	Referential Cohesion: Global Lemma Overlap	0.083
31	**DES**	**ORT**	**0.082**

offer a finer-grained assessment of students' communicative effectiveness in writing prior to beginning their academic programs [20] and have been documented before and associated with "discrete traits" of academic writing quality [29, 43]. Given ongoing concerns about disparities in placement outcomes across demographic lines, integrating these features with the sociodemographic variables identified for this study (gender and race) allows for a more comprehensive understanding of how writing traits intersect with placement decisions.

3.4 Statistical Tests

Three commonly used statistical tests, validated by the literature, were employed in this study: a one-way ANOVA, Tukey's HSD, Chi-square test of independence.

The one-way ANOVA test was selected to determine whether statistically significant differences existed in linguistic feature means across groups with more than two categories, specifically gender and race, as supported in the literature [36]. For the interpretation of the p-values, we used the following thresholds, commonly mentioned in the literature [38]:

$p < 0.01$ indicates highly significant group differences;
$0.01 \leq p < 0.05$ indicates statistically significant group differences;
$0.05 \leq p < 0.10$ indicates marginally significant group differences;
$p \geq 0.10$ indicates not significant group differences

While a multi-way ANOVA could have been used to test for interaction effects between gender and race, it was not applied here due to sample size limitations across intersecting subgroups (race $\times$ gender). To meet the assumptions of normality and homogeneity of

variances, which are required for valid interaction testing, we opted for separate one-way ANOVAs to analyze each variable independently.

Following the ANOVA calculations, we conducted Tukey's HSD tests to perform pairwise comparisons. This post-hoc analysis identified which specific groups differed significantly and gauged the magnitude of those differences. Tukey's HSD has also been used in prior writing studies examining complexity and accuracy in student writing proficiency [4].

Finally, to examine potential associations between the demographic variables in this study and placement levels, both ACCUPLACER and human assigned, we used the Chi-square test of independence, appropriate for the categorical nature of placement outcomes.

4 Correlating Sociodemographics and Linguistic Features

Correlation with Gender

To address the second research question, namely, to what extent gender and race correlate with the use of linguistic features in student writing, we applied a one-way ANOVA test to the 210 essays completed by native English speakers. The linguistic features described in Section 3.3 were analyzed against the available demographic data for this subset.

Through a one-way ANOVA test, the means of the 40 continuous (ratios) linguistic variables were compared across three gender groups (male, female, other/undisclosed). In preparation for the calculations, we adopted the conventional $\alpha = 0.05$ threshold. The degrees of freedom between groups ($\mathrm{df_B}$) were calculated using: $\mathrm{df}_B = k - 1$, yielding $\mathrm{df}_B = 2$ given the three gender groups. For the degrees of freedom within groups ($\mathrm{df_W}$), the following formula was used: $\mathrm{df}_W = N - k$. While we have 210 individual data points, 3 groups' means are calculated; therefore, $\mathrm{df}_W = 207$, which is the number of degrees of freedom that remain for estimating variability within the groups. For subsequent experiments the same formulas were employed. All statistical calculations in this study were performed using Python.

When the ANOVA calculations were performed, 36 out of the 40 linguistic features showed no statistically significant differences across the three gender groups, as summarized in Table 6. Of the remaining four features, *Positive connectives incidence* (in bold), which captures words that aid in discourse flow and cohesion (e.g., *however*, *similarly*), showed a p-value indicating a significant difference across gender groups ($p = 0.007$), a textual characteristic previously examined in writing assessment research [37]. The other three features (in italics), *Lexical Richness: HDD (excluding punctuation and numbers)* (p-value = 0.051); *Lexical Richness: Type Token Ratio (STTR NGSLeasy Nouns)* (p-value = 0.094); and *Number of POS Feature: Existential there Types* (p-value = 0.064), indicating marginally significant group variation, which warrants further investigation.

Correlation with Race

Following the same statistical approach described in Section 4, we examined whether differences emerged in the use of all 40 linguistic features across four racial groups (American Indian, Black or African American, White, Other/Undisclosed). This time $\mathrm{df}_B = 3$ given the four racial groups. While we still have 210 individual data points, 4 groups' means are calculated; therefore, $\mathrm{df}_W = 206$.

As shown in Table 7, 31 features showed no meaningful differences across racial groups (compared to 36 features in the gender-based analysis).

The remaining nine features, particularly those connected to syntax complexity, discourse markers, and lexical precision, revealed differences among racial groups. Out of the nine,

Table 6 ANOVA summary of linguistic features by gender group (M = male; F = female; O = other/undisclosed). Significance levels are indicated as follows: S = significant, M = marginally significant, N = not significant.

Features	Mean (M)	Mean (F)	Mean (O)	Grand Mean	F-Stat.	*p*	Interp.
Positive connectives incidence	79.792	91.356	92.644	87.863	5.020	**0.007**	**S**
Lexical Richness:							
HDD (excluding punctuation and numbers)	0.752	0.809	0.830	0.792	3.013	*0.051*	*M*
Lexical Richness: Type Token Ratio							
(STTR NGSLeasy Nouns)	64.155	81.123	74.464	75.793	2.393	*0.094*	*M*
Number of POS Feature: Existential there Types	0.656	0.695	1.400	0.700	2.791	*0.064*	*M*
Paragraph length, number of sentences in a paragraph, mean	16.766	18.184	15.800	17.695	0.581	0.560	N
LSA overlap, adjacent sentences, mean	0.192	0.195	0.157	0.193	0.442	0.644	N
LSA given/new, sentences, mean	0.292	0.305	0.271	0.301	1.946	0.145	N
LSA given/new, sentences, standard deviation	0.138	0.135	0.114	0.135	1.042	0.355	N
Lexical diversity, MTLD, all words	74.905	71.166	79.811	72.512	0.884	0.415	N
Lexical diversity, VOCD, all words	76.060	75.993	79.409	76.095	0.026	0.975	N
LSA verb overlap	0.095	0.099	0.089	0.098	0.619	0.540	N
WordNet verb overlap	0.516	0.524	0.498	0.521	0.258	0.773	N
Lexical Sophistication Feature: SUBTLEX Log. Word Freq. (AW Type)	4.155	4.227	4.106	4.202	1.678	0.189	N
Lexical Sophistication: Easy noun tokens (NGSL)	38.188	43.660	41.800	41.948	1.391	0.251	N
Lexical Sophistication: Easy noun types (NGSL)	23.578	24.858	24.600	24.462	0.253	0.777	N
Number of POS Feature: Existential there Tokens	1.156	1.227	1.600	1.214	0.198	0.820	N
Number of POS Feature: Plural noun Tokens	19.453	19.660	20.200	19.610	0.012	0.988	N
Number of POS Feature: Plural noun Types	13.391	12.695	14.200	12.943	0.246	0.782	N
Number of POS Feature: Preposition Types	16.094	16.326	17.400	16.281	0.137	0.872	N
Number of POS Feature: Verbs in past participle form Types	5.328	5.184	8.200	5.300	1.401	0.249	N
Number of Syntactic Constituents: Complex Noun Phrase	38.813	39.348	45.400	39.329	0.259	0.772	N
Number of Syntactic Constituents: Prenominal Noun Modifier	20.453	19.106	21.400	19.571	0.340	0.712	N
Number of Tokens with More Than 2 Syllables	94.875	93.780	94.600	94.133	0.011	0.989	N
Number of Unique Words	104.922	101.567	113.600	102.876	0.291	0.748	N
Number of Word Types (including Punctuation and Numbers)	163.219	166.014	174.800	165.371	0.092	0.913	N
Number of Word Types with More Than 2 Syllables	71.359	67.532	76.200	68.905	0.352	0.704	N
POS Density Feature: Existential There	0.003	0.003	0.004	0.003	0.125	0.883	N
POS Density Feature: Possessive Ending	0.002	0.001	0.003	0.002	0.568	0.568	N
Referential Cohesion: Global Lemma Overlap	0.609	0.726	0.770	0.691	1.125	0.327	N
MWE	0.072	0.071	0.061	0.071	0.121	0.886	N
VDISAGREE	0.002	0.003	0.003	0.003	0.683	0.506	N
ORT	0.051	0.046	0.030	0.047	0.855	0.427	N
PUNCT	0.042	0.043	0.041	0.043	0.044	0.957	N
WORDOMIT	0.006	0.008	0.005	0.007	0.821	0.442	N
PRECISION	0.011	0.011	0.006	0.011	0.436	0.648	N
WORDREPT	0.002	0.002	0.003	0.002	0.538	0.585	N
REASON	0.005	0.006	0.005	0.005	0.196	0.822	N
WORDSPLIT	0.002	0.002	0.001	0.002	0.306	0.737	N
EXAMPLE	0.002	0.003	0.002	0.002	0.154	0.858	N
ALTERN	0.001	0.002	0.001	0.001	0.084	0.920	N

four of them (bolded), namely, *Positive Connectives Incidence*, *PUNCT*, *WORDOMIT*, and *PRECISION*, yielded *p*-values that indicated significant differences across race groups. Two features (also bolded), *Lexical Sophistication (SUBTLEX Logarithmic Word Frequency)* and *Number of Syntactic Constituents: Prenominal Noun Modifier*, produced *p*-values suggesting statistically significant group differences. The remaining three (in italics), *WordNet Verb Overlap*, *ORT*, and *REASON*, were considered marginally significant, indicating that how these groups exhibit (or not) these linguistic features in their written productions requires further investigation.

Five of these features belong to the DES feature set, suggesting that racial group differences may be more pronounced when assessment focuses on more targeted, humanly-devised indicators of developmental writing. Notably, the *Positive Connectives Incidence* feature emerged as significant in both the gender and race analyses, potentially highlighting its importance in capturing variation in students' writing patterns. The differences here analyzed were not uniformly distributed across all features, indicating that group identity

Table 7 ANOVA summary of linguistic features by race group (1 = American Indian; 2 = Black or African American; 3 = White; 5 = Other/Undisclosed). Significance levels are indicated as follows: S = significant, St = statistically significant; M = marginally significant, N = not significant.

Features	Mean				Grand Mean	F-Stat.	*p*	Interp.
	(1)	(2)	(3)	(5)				
PUNCT	0.037	0.056	0.038	0.091	0.043	6.674	**0.000**	**S**
WORDOMIT	0.006	0.013	0.006	0.011	0.007	8.796	**0.000**	**S**
PRECISION	0.009	0.018	0.009	0.023	0.011	5.871	**0.001**	**S**
Positive connectives incidence	78.282	96.936	86.524	108.571	87.863	4.439	**0.005**	**St**
Lexical Sophistication Feature:								
SUBTLEX Log. Word Freq. (AW Type)	4.169	4.301	4.167	4.634	4.202	3.634	**0.014**	**St**
Number of Syntactic Constituents:								
Prenominal Noun Modifier	20.546	15.706	20.992	7.000	19.571	2.913	**0.035**	**St**
WordNet verb overlap	0.502	0.553	0.513	0.576	0.521	2.200	*0.089*	*M*
ORT	0.045	0.059	0.042	0.091	0.047	2.553	*0.057*	*M*
REASON	0.005	0.008	0.005	0.000	0.005	2.509	*0.060*	*M*
Paragraph length, number of sentences in a paragraph, mean	17.576	16.353	18.368	6.000	17.695	1.038	0.377	N
LSA overlap, adjacent sentences, mean	0.197	0.213	0.183	0.283	0.193	1.746	0.159	N
LSA given/new, sentences, mean	0.304	0.300	0.300	0.314	0.301	0.061	0.980	N
LSA given/new, sentences, standard deviation	0.141	0.141	0.132	0.165	0.135	1.385	0.249	N
Lexical diversity, MTLD, all words	73.130	68.164	74.002	87.500	72.512	0.981	0.403	N
Lexical diversity, VOCD, all words	75.327	70.684	78.469	80.586	76.095	0.687	0.561	N
LSA verb overlap	0.094	0.100	0.098	0.086	0.098	0.264	0.851	N
Lexical Richness: HDD (excluding punctuation and numbers)	0.773	0.783	0.801	0.825	0.792	0.354	0.787	N
Lexical Richness: Type Token Ratio								
(STTR NGSLeasy Nouns)	70.466	74.924	78.026	16.900	75.793	0.625	0.599	N
Lexical Sophistication: Easy noun tokens (NGSL)	40.000	39.765	43.584	13.000	41.948	1.081	0.358	N
Lexical Sophistication: Easy noun types (NGSL)	23.303	22.726	25.592	10.000	24.462	1.341	0.262	N
Number of POS Feature: Existential there Tokens	0.909	1.275	1.280	0.000	1.214	0.713	0.546	N
Number of POS Feature: Existential there Types	0.576	0.647	0.760	0.000	0.700	1.133	0.337	N
Number of POS Feature: Plural noun Tokens	19.455	17.412	20.632	9.000	19.610	1.112	0.345	N
Number of POS Feature: Plural noun Types	13.970	11.059	13.496	6.000	12.943	1.719	0.164	N
Number of POS Feature: Preposition Types	15.758	15.863	16.624	12.000	16.281	0.532	0.661	N
Number of POS Feature:								
Verbs in past participle form Types	5.121	4.412	5.736	2.000	5.300	1.620	0.186	N
Number of Syntactic Constituents: Complex Noun Phrase	37.697	36.059	41.256	19.000	39.329	1.309	0.272	N
Number of Tokens with More Than 2 Syllables	92.909	85.392	98.496	35.000	94.133	1.316	0.270	N
Number of Unique Words	102.546	93.961	106.920	63.000	102.876	1.395	0.246	N
Number of Word Types								
(including Punctuation and Numbers)	160.727	155.137	171.280	102.000	165.371	1.113	0.345	N
Number of Word Types with More Than 2 Syllables	69.515	59.686	72.840	27.000	68.905	2.101	0.101	N
POS Density Feature: Existential There	0.003	0.003	0.004	0.000	0.003	0.516	0.672	N
POS Density Feature: Possessive Ending	0.002	0.002	0.001	0.000	0.002	0.052	0.984	N
Referential Cohesion: Global Lemma Overlap	0.605	0.689	0.719	0.200	0.691	0.680	0.565	N
MWE	0.069	0.078	0.069	0.120	0.071	0.844	0.471	N
VDISAGREE	0.003	0.003	0.002	0.011	0.003	0.967	0.409	N
WORDREPT	0.001	0.003	0.002	0.000	0.002	0.754	0.521	N
WORDSPLIT	0.002	0.002	0.002	0.000	0.002	0.071	0.976	N
EXAMPLE	0.002	0.002	0.003	0.000	0.002	0.617	0.605	N
ALTERN	0.002	0.002	0.001	0.000	0.001	0.368	0.776	N

may influence only specific aspects of writing.

Beyond the One-way ANOVA

While the one-way ANOVA identified overall group differences by gender and race, Tukey's HSD post-hoc test was used to determine which specific groups differed from one another.

For the gender analysis, we zeroed in on the statistical differences highlighted in Section 4[11]. The sample analyzed consisted of 141 females (F) and 64 males (M), for a total of 205 participants. The "Other/Undisclosed" category, which had only 5 participants, was excluded in an attempt to make comparisons involving the other two groups as statistically stable as possible. Consequently, for the Tukey's HSD test, the following values were used: $k = 2$ (groups: Male, Female); $\mathrm{df}_W = 203$; $\alpha = 0.05$; $q_{\text{critical}} = 2.788$.

[11] For brevity, Tukey's HSD tests were only conducted on this feature set; other features, including those identified for racial groups, will be analyzed in future work.

Table 8 Tukey's HSD test results for selected features for all gender groups: M = males; F = females. Significance levels are indicated as follows: N = not significant.

Feature	Mean Dif. (F-M)	Stand. Err.	q-Stat	Sign.
Positive connectives incidence	11.564	5.161	2.241	N
Lexical Richness: HDD (excluding punctuation and numbers)	0.058	0.033	1.736	N
Lexical Richness: Type Token Ratio (STTR NGSLeasy Nouns)	16.968	10.970	1.547	N
Number of POS Feature: Existential there Types	0.039	0.023	1.671	N

From Table 8, we see that all pairwise comparisons by gender were not statistically significant at $\alpha = 0.05$, as the calculated q-statistics for all four linguistic features examined fell below the critical value of $q_{\text{critical}} = 2.788$. Although some mean differences, particularly for *Positive Connectives Incidence* were notable, they did not exceed the threshold for significance. Consequently, we cannot reject the null hypothesis, which posits no meaningful differences in the use of linguistic features across gender groups. These findings should be interpreted with caution, as the limited sample size and unequal group distributions could limit the statistical power to detect more subtle effects.

Table 9 Tukey's HSD test results for selected features for all race groups: 1 = American Indian; 2 = Black or African American; 3 = White.

Feature	Mean Diff			Stand. Error	q-Stat		
	(1–2)	(1–3)	(2–3)		(1–2)	(1–3)	(2–3)
PUNCT	0.019	0.001	0.018	0.006	**3.247**	0.174	**3.074**
WORDOMIT	0.007	0.000	0.007	0.002	**3.606**	0.052	**3.658**
PRECISION	0.009	0.001	0.008	0.003	**3.065**	0.204	**2.860**
Positive connectives incidence	18.654	8.243	10.411	6.275	**2.973**	1.314	1.659
Lexical Sophistication Feature:							
SUBTLEX Log Word Freq (AW Type)	0.132	0.002	0.134	0.057	2.319	0.032	2.350
Number of Syntactic Constituents:							
Prenominal Noun Modifier	4.840	0.447	5.286	2.429	1.992	0.184	2.176
WordNet verb overlap	0.050	0.011	0.040	0.025	1.990	0.419	1.570
ORT	0.013	0.003	0.017	0.008	1.717	0.413	2.131
REASON	0.002	0.001	0.003	0.001	1.614	0.538	2.152

We then turned to the post-hoc analysis for the race groups. For Tukey's HSD test, the following values were used: $k = 3$ [race groups: American Indian (1), Black or African American (2), and White (3)]. The Other/Undisclosed race category previously mentioned included only one student. Due to this small count, this category was excluded from the statistical analysis in an attempt to preserve the reliability of the test results. For $\text{df}_W = 208$; $\alpha = 0.05$; $q_{\text{critical}} = 2.788$.

As shown in Table 9, statistically significant differences (in bold) were found for *WORDOMIT*, *PUNCT*, and *PRECISION* between Groups 1 and 2 as well as Groups 2 and 3. A borderline yet statistically significant difference in *Positive Connectives Incidence* was also observed between Groups 1 and 2 only. Based on the individual means already reported in Table 7, students who identified as Black or African American tended to exhibit a higher incidence of omitted or left-out parts of speech compared to both American Indian and White students. These omissions may affect the overall coherence of their writing. In contrast, American Indian and White students showed nearly identical omission ratios.

Students identifying as Black or African American also used punctuation (e.g., for joining clauses) less frequently than their peers, and exhibited a higher rate of imprecise word usage to describe a concept, as captured by the *PRECISION* feature. On the contrary, Black or African American students showed a higher incidence of *Positive Connectives* than American Indian students, suggesting greater use of linking expressions such as *however* or *similarly*.

These differences may reflect dialectal or regional variation rather than deficits in academic ability [34]. However, given that placement decisions are tied to institutional standards that may not fully account for this linguistic diversity, it is essential that pedagogical interventions not be seen as merely corrective measures but as opportunities to support students in expanding their register and linguistic skills to better navigate academic expectations.

It is equally important to acknowledge that these differences were identified from samples produced at a single point in time, and that external factors, such as test anxiety or unfamiliarity with the writing prompt, may have also influenced students' performance.

5 Assessing Level Assignment with Demographics

Assesment with Gender

To evaluate the possible association between gender and ACCUPLACER placement levels, in response to our third research question, a Chi-square test of independence was performed as the final step in this analysis. Table 10 presents the observed counts, expected counts under the null hypothesis of independence, and the individual Chi-square contributions $(O-E)^2/E$ for each cell.

Table 10 Observed vs. Expected Counts for ACCUPLACER Placement by Gender with Chi-square Contributions.

Placement Level	Gender	Observed	Expected	$(O - E)^2 / E$
DevEd Level 1	Male	21	16.859	1.017
	Female	33	37.141	0.462
DevEd Level 2	Male	20	24.039	0.679
	Female	57	52.961	0.308
College Level	Male	23	23.102	0.000
	Female	51	50.898	0.000
Total	-	205	205	2.467

The expected counts were calculated based on the assumption that gender and placement level are independent, and the $(O-E)^2/E$ values represent each cell's contribution to the overall Chi-square statistic. The total sum of these contributions yielded the total Chi-square value ($\chi^2 = 2.467$). To assess statistical significance, we compared the calculated χ^2 value to the critical Chi-square value ($\chi^2_{\text{critical}} = 5.991$), determined for $df = 2$ and $\alpha = 0.05$. Alternatively, the associated p-value ($p = 0.291$) was considered. Since $\chi^2 < \chi^2_{\text{critical}}$ and $p > 0.05$, we cannot reject the null hypothesis and conclude that there is no statistically significant association between gender and the automatically assigned skill levels via ACCUPLACER.

The same procedure was applied using the human-assigned skill levels, with identical degrees of freedom and significance thresholds. Because human raters assigned different placement levels than ACCUPLACER, the distribution of texts by levels varies.

The expected counts, as included in Table 11, were calculated based on the same assumption that gender and placement level are independent, and the $(O-E)^2/E$ values represent each cell's contribution to the overall Chi-square statistic. The total sum of these contributions yielded the total Chi-square value ($\chi^2 = 0.415$). The critical Chi-square value was the same ($\chi^2_{\text{critical}} = 5.991$). Alternatively, the associated p-value ($p = 0.812$) was considered. Since $\chi^2 < \chi^2_{\text{critical}}$ and $p > 0.05$, we also conclude that there is no statistically significant association between gender and the human-assigned placement levels.

Table 11 Observed vs. Expected Counts for Human Placement by Gender with Chi-square Contributions.

Placement Level	Gender	Observed	Expected	$(O - E)^2 / E$
DevEd Level 1	Male	14	13.112	0.060
	Female	28	28.888	0.027
DevEd Level 2	Male	36	38.088	0.114
	Female	86	83.912	0.051
College Level	Male	14	12.8	0.112
	Female	27	28.2	0.051
Total	-	205	205	0.415

Assessment with Race

We also examined whether there was an association first between racial groups and the classification by skill level produced by ACCUPLACER, and, second, with the human ratings. The distribution of placement levels across racial groups, along with their corresponding observed and expected counts and Chi-square contributions, is presented in Table 12.

Table 12 Observed vs. Expected Counts for (ACCUPLACER) Placement by Race with Chi-square Contributions.

Placement Level	Race	Observed	Expected	$(O - E)^2 / E$
DevEd Level 1	American Indian	10	8.526	0.255
	Black or African American	14	13.177	0.051
	White	30	32.297	0.163
DevEd Level 2	American Indian	10	12.474	0.491
	Black or African American	25	19.278	1.699
	White	44	47.249	0.223
College Level	American Indian	13	12.000	0.083
	Black or African American	12	18.546	2.310
	White	51	45.455	0.677
Total	-	209	209	5.952

The degrees of freedom were calculated based on 3 three levels (DevEd Level 1, DevEd Level 2, College Level) and 3 racial groups (American Indian, Black or African American, and White). The computed Chi-square statistic yielded a value of $\chi^2 = 5.952$, which does not exceed the critical value of 9.488 for 4 degrees of freedom at the $\alpha = 0.05$ level. The associated p-value of $p = 0.203$ is above the conventional threshold for statistical significance. Therefore, we cannot reject the null hypothesis, indicating no statistically significant association between race and ACCUPLACER placement level within this sample.

Finally, the assessment was repeated but with the humanly-assigned skill level. Table 13 summarizes the observed and expected counts and Chi-square contributions.

The computed Chi-square statistic yielded a value of $\chi^2 = 9.588$, which slightly exceeds the critical value of 9.488 for 4 degrees of freedom at the $\alpha = 0.05$ level. The associated p-value of $p = 0.048$ falls just below the conventional threshold for statistical significance. While this suggests a possible association between race and human-assigned placement levels, even though raters did not have access to demographic data, the result is borderline and should be interpreted cautiously, warranting further investigation.

Upon analyzing each individual Chi-square contribution, we focused on values with the greatest discrepancies between the observed and expected counts within the race. This approach enabled us to identify, within the context of this sample, whether certain racial

Table 13 Observed vs. Expected Counts for (Human) Placement by Race with Chi-square Contributions.

Placement Level	Race	Observed	Expected	(O - E)² / E
DevEd Level 1	American Indian	4	6.914	1.228
	Black or African American	*17*	*10.686*	3.731
	White	22	26.191	0.671
DevEd Level 2	American Indian	20	19.486	0.014
	Black or African American	29	30.114	0.041
	White	*75*	*73.810*	0.019
College Level	American Indian	9	6.600	0.873
	Black or African American	*5*	*10.200*	2.651
	White	28	25.000	0.360
Total	-	209	209	9.588

groups were overrepresented or underrepresented at specific placement levels. A higher observed count relative to the expected suggests overrepresentation, with a lower observed count indicating underrepresentation (italicized values).

In the DevEd Level 1 category, students who self-identified as Black or African American appeared to be more frequently than expected placed within this level, which may help explain their comparatively lower representation at the College Level. In contrast, students who identified as White were more frequently placed in DevEd Level 2 and College Level, indicating a higher-than-expected concentration in those categories. The variations between the observed and expected contributions for American Indians, within each placement level, were not as pronounced as those of the other groups.

6 Conclusions

This study investigated the intersection of sociodemographic characteristics, linguistic features, and writing placement outcomes in a higher education context. We analyzed a corpus of 210 anonymized writing samples from native English speakers (L1), representative of the institution's demographic composition. Using 40 top-ranked linguistic features drawn from Coh-Metrix, CTAP, and Developmental Education-Specific (DES) feature sets, we assessed potential disparities across gender and race using three statistical tests: one-way ANOVA, Tukey's HSD, and Chi-square.

The results we obtained provided evidence that not all linguistic features are equally sensitive to demographic variation. Gender, in particular, does not appear to have an influence on the linguistic features measured. This finding supports the idea that, at least for the features examined, variation in writing may not be primarily driven by gender.

Among racial groups, in contrast, a total of 9 features, particularly those tied to syntax complexity, discourse markers, and lexical precision, revealed differences. These features were mostly found within the human-devised set (DES) and aligned with the focus of DevEd courses. While not designed to directly measure academic ability, these DES features offer valuable diagnostic insight into students' current writing performance at the onset of college and align with DevEd's mission of identifying support needs for student success.

Comparing automated and human-assigned placement classifications, no significant associations were found with gender. However, a borderline significant association between race and human-assigned placement levels ($p = 0.048$) raises important questions about raters' sensitivity to linguistic features that intersect with race. This finding highlights the need for continued efforts to ensure placement fairness and address potential bias. Because

human evaluation remains essential for detecting nuanced linguistic patterns that automated systems may overlook, our findings advocate for a combined approach that leverages the strengths of human judgment and automated analysis to potentially: (i) enhance placement accuracy and educational outcomes; (ii) guide targeted instruction; and (iii) support the development of more equitable and responsive placement tools.

References

1 Hervé Abdi and Lynne J Williams. Tukey's Honestly Significant Difference (HSD) test. *Encyclopedia of Research Design*, 3(1):1–5, 2010.

2 Lisa R Arnold, Lei Jiang, and Holly Hassel. After implementation: Assessing student self-placement in college writing programs. *Journal of Writing Assessment*, 17(1), 2024.

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

4 Jessie S Barrot and Joan Y Agdeppa. Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, 47:100510, 2021.

5 Susan Bickerstaff, Katie Beal, Julia Raufman, Erika B Lewy, and Austin Slaughter. Five principles for reforming Developmental Education: A review of the evidence. *Center for the Analysis of Postsecondary Readiness*, pages 1–8, 2022.

6 Carolina Castillo and Natalia Ávila Reyes. Students' sociodemographic characteristics and writing performance: A systematic literature review. *Reading and Writing*, pages 1–37, 2025.

7 Xiaobin Chen and Detmar Meurers. 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, December 2016. The COLING 2016 Organizing Committee. URL: `https://aclanthology.org/W16-4113`.

8 Miguel Da Corte and Jorge Baptista. Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3046–3056, Torino, Italia, May 2024. ELRA and ICCL. URL: `https://aclanthology.org/2024.lrec-main.272/`.

9 Miguel Da Corte and Jorge Baptista. Enhancing writing proficiency classification in Developmental Education: The quest for accuracy. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, Torino, Italia, May 2024. ELRA and ICCL. URL: `https://aclanthology.org/2024.lrec-main.542/`.

10 Miguel Da Corte and Jorge Baptista. Leveraging NLP and machine learning for English (l1) writing assessment in developmental education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140, 2024. `doi:10.5220/0012740500003693`.

11 Miguel Da Corte and Jorge Baptista. Refining English writing proficiency assessment and placement in Developmental Education using NLP tools and Machine Learning. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 288–303. INSTICC, SciTePress, 2025. `doi:10.5220/0013351500003932`.

12 Miguel Da Corte and Jorge Baptista. Toward consistency in writing proficiency assessment: Mitigating classification variability in Developmental Education. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 139–150. INSTICC, SciTePress, 2025. `doi:10.5220/0013353900003932`.

13 Jane Denison-Furness, Stacey Lee Donohue, Annemarie Hamlin, and Tony Russell. 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, 2022. doi:10.37514/PRA-B.2022.1565.2.04.

14 Martin East and David Slomp. The ethical turn in writing assessment: How far have we come, and where do we still need to go? *Language Teaching*, 57(2):262–273, 2024.

15 Nikki Edgecombe and Michael Weiss. Promoting equity in Developmental Education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1, 2024.

16 Elizabeth Ganga and Amy Mazzariello. Modernzing college course placement by using multiple measures. *Education Commission of the States*, pages 1–9, 2019. URL: https://postsecondaryreadiness.org/wp-content/uploads/2019/03/Modernizing_College_Course_Placement_by_Using_Multiple_Measures_Final.pdf.

17 Sandra Götz and Sylviane Granger. Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1):1–38, 2024.

18 Sarah Hirsch, Kenny Smith, and Madeleine Sorapure. Collaborative writing placement: Partnering with students in the placement process. *Journal of Writing Assessment*, 17(2), 2024.

19 Darin L Jensen and Joanne Baird Giordano. Afterword. Placement, equity, and the promise of democratic open-access education. *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education*, pages 279–86, 2022.

20 Young-Suk Grace Kim, Christopher Schatschneider, Jeanne Wanzek, Brandy Gatlin, and Stephanie Al Otaiba. Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and writing*, 30:1287–1310, 2017.

21 Holly Kosiewicz, Cristián Morales, and Kalena E. Cortes. The "missing English learner"' in higher education: How identification, assessment, and placement shape the educational outcomes of English learners in community colleges. In *Higher Education: Handbook of Theory and Research: Volume 39*, pages 1–55. Springer, 2023.

22 Josh Lederman. Validity and racial justice in educational assessment. *Applied Measurement in Education*, 36(3):242–254, 2023. doi:10.1080/08957347.2023.2214654.

23 Bruce W Lee and Jason Hyung-Jong Lee. Prompt-based learning for text readability assessment. In *In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19. Toronto, Canada. Association forComputational Linguistics., 2023.

24 Stephanie Link and Svetlana Koltovskaia. *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham, 2023. doi:10.1007/978-3-031-36033-6_21.

25 Susan Lyons, Maria Elena Oliveri, and Mya Poe. A framework for enacting equity aims in assessment use: A justice-oriented approach. In *Culturally Responsive Assessment in Classrooms and Large-Scale Contexts*, pages 88–105. Routledge, 2025.

26 Ross Markle. Redesigning course placement in service of guided pathways. *Educational Considerations*, 50(2):8, 2025.

27 Michael Matta and Narmene Hamsho and. Consequences of response formats on racial and ethnic bias and fairness in writing assessments. *School Psychology Review*, 0(0):1–14, 2025. doi:10.1080/2372966X.2025.2462984.

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

29 Neil Murray and Gerard Sharpling and. What traits do academics value in student writing? Insights from a psychometric approach. *Assessment & Evaluation in Higher Education*, 44(3):489–500, 2019. doi:10.1080/02602938.2018.1521372.

30 Jessica Nastal. Beyond tradition: Writing placement, fairness, and success at a two-year college. *Journal of Writing Assessment*, 12(1), 2019.

31 Jessica Nastal and Kris Messer. Afterword: Finding the right note in writing placement. *Journal of Writing Assessment*, 18(1), 2025.

32 Mari Nygård and Anne Kathrine Hundal. Features of grammatical writing competence among early writers in a Norwegian school context. *Languages*, 9(1):29, 2024.

33 María Elena Oliveri, René Lawless, and Robert J. Mislevy. Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, 19(3):270–300, 2019. doi:10.1080/15305058.2018.1543308.

34 Ramona T Pittman, Lynette O'Neal, Kimberly Wright, and Brittany R White. Elevating students' oral and written language: Empowering African American students through language. *Education Sciences*, 14(11):1191, 2024.

35 Mya Poe, Jessica Nastal, and Norbert Elliot. Reflection. An admitted student is a qualified student: A roadmap for writing placement in the two-year college. *Journal of Writing Assessment*, 12(1), 2019.

36 Jennifer Reid, Mahshid Ahmadian, D. Jennings, Anathea Abad Pepperl, and et.al. Saying it aloud: Inclusive teaching statements impact on sense of belonging and engagement. *Journal of College Science Teaching*, 0(0):1–14, 2025. doi:10.1080/0047231X.2025.2487437.

37 Rachael Ruegg. Assessment of written assignments in first-year Humanities and Social Sciences courses: Textual features of academic writing. *Assessment in Education: Principles, Policy & Practice*, 32(1):60–76, 2025. doi:10.1080/0969594X.2025.2467672.

38 Padam Singh. P value, statistical significance and clinical significance. *Journal of Clinical and Preventive Cardiology*, 2(4):202–204, 2013.

39 J Suárez-Álvarez, M Oliveri, A Zenisky, and SG Sireci. Current assessment needs in adult education and workforce development: Summary report (center for educational assessment report no. 998). *Center for Educational Assessment*, 2023. URL: https://createadultskills.org/system/files/ASAP%20Brief%201_Nov%202023_Needs%20Assessment.pdf.

40 Meghan Sweeney and Crystal Colombini. (Re) placing personalis: A study of placement reform and self-construction in mission-driven contexts. *Journal of Writing Assessment*, 17(1), 2024.

41 The College Board. ACCUPLACER Program Manual. (online), 2022. URL: https://secure-media.collegeboard.org/digitalServices/pdf/accuplacer/accuplacer-program-manual.pdf.

42 Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, 2022. URL: https://aclanthology.org/2022.lrec-1.130.

43 Sachiko Yasuda. What does it mean to construct an argument in academic writing? a synthesis of English for general academic purposes and English for specific academic purposes perspectives. *Journal of English for Academic Purposes*, 66:101307, 2023.

Section 4
Exploring Next-Generation Writing Placement: Leveraging LLMs for DevEd

Section 4, titled *Exploring Next-Generation Writing Placement: Leveraging LLMs for DevEd*, serves as the final section of this thesis and introduces exploratory work aimed at rethinking placement practices within the context of community colleges in the U.S. Persistent challenges with overplacement and underplacement with current automated systems have led institutions to pursue more transparent, equitable, and reliable alternative tools, particularly Large Language Models (LLMs), among them (Seßler et al. (2025)). This section evaluates whether, and how, models such as OpenAI ChatGPT, Claude, and Gemini might support writing assessment and placement in a DevEd context, and responds to the fourth research question (RQ4) of this thesis: *What is the potential of LLMs to serve as complementary tools to existing automated placement systems*?

Several studies support this direction, highlighting the potential of LLMs to provide automated writing feedback and instructional support to students (Adiguzel et al. (2023); Nikolic et al. (2024)). However, questions remain about the reliability of these models in assessing writing proficiency and aligning with institutional standards of proficiency. Further research is needed to examine how human-annotated corpora can be integrated into LLMs' writing assessment logic to enhance placement accuracy and reliability. Such integration could help provide better-informed placement decisions with more linguistically grounded descriptors (Miaschi et al. (2024)) and guide incoming students more effectively through the lens of language proficiency and college readiness.

The paper featured in this section, *From Prediction to Precision: Leveraging LLMs for Equitable and Data-Driven Writing Placement in Developmental Education* (Da Corte and Baptista (2025b)), zeroes in on the evaluation of LLMs as potential complementary tools for placement. This study builds on the 300 essays that formed the primary corpus described in the Study's Participants and Corpus section of the Introduction. Using the same random sample selection technique employed throughout this investigation, the corpus was expanded to include 150 additional texts, which were automatically classified by ACCUPLACER as college-level and manually annotated and assessed by the two trained raters supporting this study[51]. The final dataset now comprises 450 English-language essays written by college-intending students, evenly distributed across three levels: DevEd Level 1, DevEd Level 2, and College-Level (150 texts each). This expanded and balanced corpus enabled a more comprehensive evaluation of writing proficiency, capturing a broader range of performance across both developmental and college-level writing.

[51] **Appendix G** presents an annotated College-level sample essay response from the expanded corpus, with its automatic classification score issued by ACCUPLACER.

With this expanded scope, the goal was to assess how precisely LLMs[52], namely Claude 3.5 Sonnet, Gemini 2.0 Flash, ChatGPT-3.5-turbo, and ChatGPT-4o, could (i) classify texts as proficient or non-proficient (College level versus Non-college level), and (ii) generate rationales behind their classification decisions. Their classifications were compared to those produced by ACCUPLACER and by human raters. The latter served as the gold standard benchmark. All four models were tested using their default hyperparameters and were selected based on a combination of practical and performance-based criteria, including API availability, reasoning capabilities, and cost-efficiency. Recent research has also cited these models as representative of emerging best practices in educational LLM deployment (Wang et al. (2024b)).

For the assessment itself, framed as a classification task, two classification strategies were employed, with precision selected as the primary evaluation metric, given its critical role in minimizing misplacement in DevEd contexts: (1) a one-step classification, which directly assigned all 450 essays to one of three levels in a single pass: DevEd Level 1, DevEd Level 2, or College-level; and (2) a two-step classification, which first distinguished between College-level and DevEd texts, and then further classified the DevEd subset into Level 1 or Level 2. The two-step strategy was designed to reduce complexity and improve precision when separating adjacent developmental levels, a strategy supported by prior research (Fields et al. (2024)). Each classification strategy included four prompting conditions: (a) Zero-shot; (b) Few-shot; (c) Enhanced; and (d) Enhanced+; all detailed in the paper. The prompts associated with each strategy were literature-informed, human-authored, and iteratively refined to reflect authentic linguistic markers of writing proficiency (Min et al. (2023); Xiao et al. (2025)), and are included in **Appendix H**.

In terms of results, Claude achieved an overall precision of $\approx$ 62% in the one-step classification strategy using the Enhanced+ prompt, outperforming ACCUPLACER's $\approx$ 58%. This precision score means that, under the given scenario, Claude accurately placed about six out of ten students. Most misclassifications were underplacements, suggesting that Claude applied a stricter threshold than human raters. The model correctly classified $\approx$ 97% of DevEd Level 1 texts but struggled with DevEd Level 2 ($\approx$ 46% accuracy) and performed slightly better at College Level ($\approx$ 61%). When assessing alignment with the gold standard (human raters), Claude's classifications under the Enhanced+ prompt showed a *strong* Pearson correlation of 0.75, compared to a slightly lower but still *substantial* correlation of 0.67 for ACCUPLACER. This difference suggests that the assessment and classification of certain LLMs may more closely approximate that of trained human assessments.

In the two-step classification strategy[53], precision scores improved for most models compared to the one-step, except for ChatGPT-4o. Gemini achieved the highest precision at $\approx$ 69%,

[52]This study was completed before the release of ChatGPT-5. Future text classification task experiments will incorporate this model to compare its precision with those presented here.

[53]The dataset with the full results from the 2-step classification experiment is publicly available through GitHub to support transparency and reproducibility in evaluating model performance. Corpus content cannot yet be released due to ongoing IRB procedures.

which is interpreted as the model accurately placing nearly seven out of ten students. Most misclassifications occurred between neighboring levels, specifically between DevEd Level 1 and DevEd Level 2, and between the College Level and DevEd Level 2. Gemini correctly classified $\approx 71\%$ of DevEd Level 1 texts, $\approx 76\%$ of DevEd Level 2, and $\approx 51\%$ of College-level texts. In this classification strategy, the alignment between LLMs and the gold standard (human raters) was also calculated. Gemini's classifications under the Enhanced+ prompt showed a *strong* Pearson correlation of 0.67, compared to a slightly lower correlation of 0.67 for ACCUPLACER. The highest correlation was, again, achieved by Claude (0.73) within the same prompting strategy.

By demonstrating that LLMs, particularly in the case of Claude and Gemini, can match or outperform traditional systems like ACCUPLACER in certain scenarios, this study introduces some considerations for rethinking placement practices in higher education. These considerations set the stage for exploring hybrid models that combine human judgment, annotated corpora, and advanced language models to better support student success from the start of their academic journey.

Reference to the paper in this Section:

Paper 9:
Da Corte, M. and Baptista, J. (2025b). **From Prediction to Precision: Leveraging LLMs for Equitable and Data-driven Writing Placement in Developmental Education**. In Baptista, J. and Barateiro, J., editors, 14th Symposium on Languages, Applications and Technologies (SLATE 2025), volume 135 of Open Access Series in Informatics (OASIcs), pages 1:1–1:18, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

From Prediction to Precision: Leveraging LLMs for Equitable and Data-Driven Writing Placement in Developmental Education

Miguel Da Corte
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Jorge Baptista
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID Lisboa, Portugal

Abstract

Accurate text classification and placement remain challenges in U.S. higher education, with traditional automated systems like ACCUPLACER functioning as "black-box" models with limited assessment transparency. This study evaluates Large Language Models (LLMs) as complementary placement tools by comparing their classification performance against a human-rated gold standard and ACCUPLACER. A 450-essay corpus was classified using Claude, Gemini, GPT-3.5-turbo, and GPT-4o across four prompting strategies: Zero-shot, Few-shot, Enhanced, and Enhanced+ (definitions with examples). Two classification approaches were tested: (i) a 1-step, 3 class classification task, distinguishing DevEd Level 1, DevEd Level 2, and College-level texts in one single run; and (ii) a 2-step classification task, first separating College vs. Non-College texts before further classifying Non-College texts into DevEd sublevels. The results show that structured prompt refinement improves the precision of LLMs' classification, with Claude Enhanced + achieving 62.22% precision (1 step) and Gemini Enhanced + reaching 69.33% (2 step), both surpassing ACCUPLACER (58.22%). Gemini and Claude also demonstrated strong correlation with human ratings, with Claude achieving the highest Pearson scores ($\rho = 0.75$; 1-step, $\rho = 0.73$; 2-step) vs. ACCUPLACER ($\rho = 0.67$). While LLMs show promise for DevEd placement, their precision remains a work in progress, highlighting the need for further refinement and safeguards to ensure ethical and equitable placement.

2012 ACM Subject Classification Social and professional topics → Adult education; Social and professional topics → Student assessment; Computing methodologies → Information extraction; Computing methodologies → Lexical semantics; Computing methodologies → Language resources; Computing methodologies → Natural language generation

Keywords and phrases Large Language Models (LLMs), Developmental Education (DevEd), writing assessment, text classification, English writing proficiency

Digital Object Identifier 10.4230/OASIcs.SLATE.2025.1

Supplementary Material *Software*: `https://gitlab.hlt.inesc-id.pt/u000803/deved`

Funding 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).

1 Introduction and Objectives

Higher education institutions play a critical role in developing students' academic skills and ensuring equitable access to educational opportunities, particularly for those whose writing and reading abilities do not yet meet college-level standards. In the United States of America, this support is provided through a foundational literacy program known as Developmental

Education (DevEd). Placement in DevEd remains a persistent challenge, with traditional tools like ACCUPLACER[1] operating as "black box" automated systems. These assessments rely on machine-scored essays, emphasizing surface-level features (e.g., *word count*, *syntactic complexity*, *cohesion*) while offering limited insight into students' true writing proficiency. As a result, misplacement is common and could hinder student success [12] and progression in an academic program. In this paper, we focus on the writing skills of college-intending students at the onset of their academic career.

To address these limitations, institutions are exploring more transparent and precise assessment and course placement methods. LLMs have emerged as promising tools for placement and literacy instruction due to their writing assistance capabilities and potential for automated feedback [1, 33]. However, their reliability in assessing writing proficiency and alignment with institutional standards requires further exploration [3].

This study examines the role of LLMs in writing assessment, placement, and feature generalization within a DevEd setting. Building on previous work [10, 14], it expands an annotated corpus of 300 essays classified as Non-college level (DevEd Level 1 and DevEd Level 2) by both ACCUPLACER and human raters, adding 150 College-Level essays that were classified through the same procedures. The inclusion of College-Level texts enables a more comprehensive evaluation of LLM classification across these three proficiency levels. All writing samples were produced in English, which is the main language of instruction at the institution under study and the primary language assessed by ACCUPLACER.

In this study, we aim to:

(i) evaluate LLMs for DevEd placement by comparing their classifications to ACCUPLACER and a human-rated gold standard;
(ii) analyze the impact of a multi-step classification strategy on LLM precision across different prompting framework; and
(iii) determine whether prompt refinement mitigates classification bias, improving consistency and equity in placement outcomes.

To achieve these objectives, we propose the following research questions:

(i) How do LLM classifications compare to ACCUPLACER and human ratings in predicting placement outcomes?
(ii) What is the impact of a multi-step classification strategy on LLM precision across different prompting frameworks?
(iii) Can prompt refinement mitigate bias in LLM classification, enhancing placement accuracy and consistency across proficiency levels?

This research contributes to data-driven, equitable, and transparent DevEd placement by integrating LLMs with human-annotated corpora. By addressing systemic barriers in classification, placement, and pedagogy that hinder student success, this study explores alternative methods to refine entrance assessment tools and support linguistically underprepared students in DevEd and beyond [15]. The findings may inform more effective learning strategies, ultimately improving outcomes for students navigating writing-intensive coursework.

Our paper is structured as follows: Section 2 reviews LLMs in writing assessment and classification. Section 3 details the corpus, experimental setup, and LLMs used. Section 4 presents and discusses findings. Section 5 summarizes key contributions and future work.

[1] `https://www.accuplacer.org/` (Last accessed: 28th July 2025; all URLs in this paper were checked on this date.)

2 Related Work

Research on the use of LLMs in writing assessment, text classification, and automated feedback has gained increasing attention from higher education institutions and their policymakers [34, 41]. Several studies have evaluated LLMs, in particular to investigate their classification precision, interpretability, and feedback generation [18, 26, 42, 48] and have positioned them as potential alternatives to automated placement systems comparable to ACCUPLACER.

Despite this growing interest, we found that ACCUPLACER and similar standardized, automated systems remain widely used, yet they rely on "black-box" algorithms that prioritize surface-level linguistic features (e.g., *word count*, *syntactic complexity*) over deeper assessments of writing proficiency [16, 23] to better understand if students can communicate effectively. Some studies estimate that 30–50% of students are misplaced due to these limitations [30], raising concerns about the validity of such assessments [35]. Moreover, the proprietary nature of systems like ACCUPLACER restrict reproducibility, limit assessment transparency, and raise ethical research concerns [43].

While LLMs are also proprietary and subject to frequent updates, their integration in structured, prompt-based research provides a greater degree of process transparency and interpretability. Unlike ACCUPLACER, LLM-based classification allows researchers to design and control input conditions, evaluate model-generated justifications, and iteratively refine prompts to align with pedagogical goals. Nevertheless, we recognize that true reproducibility remains a challenge, particularly as models evolve over time. Although this caveat raises valid concerns, the promise of LLMs in advancing transparent placement practices remains a compelling one, continuing to gain traction in the field. Recent work highlights how carefully constructed LLM applications can improve placement accuracy while promoting fairness in high-stakes decisions [42, 47], which is one of the central aims of our ongoing study.

LLMs have shown promise in classification tasks across diverse text domains [42], demonstrating cost-effective assessment capabilities [38, 46]. For instance, GPT-4 and Gemini-Pro have excelled in sentiment classification for complex texts, surpassing traditional models with GPT-4 achieving the highest accuracy (0.76) [47]. GPT-4 has also outperformed Gemini in grammar correction, while Gemini excelled in sentence completion and logical reasoning [3]. In zero-shot text classification across 11 datasets, GPT-4 significantly outperformed GPT-3.5 [17]. Multi-prompting and iterative refinement strategies have been identified as effective in improving classification consistency by enabling LLMs to better internalize linguistic criteria [36], specifically, human-defined standards [29, 48].

Beyond classification, LLMs are being explored for Automated Essay Scoring (AES), where studies highlight their accuracy, consistency, and generalizability [28, 40]. When provided with clear rubrics and text examples, GPT-4 has demonstrated strong performance in AES tasks [48]. A study assessing 119 placement essays with four LLMs (Claude 2, GPT-3.5, GPT-4, and PaLM 2[2]) found that GPT-4 exhibited the highest intrarater reliability, though its performance fluctuated slightly over time [34]. Another study found GPT-4o to be the most consistent AES model (ICC above 0.7), while older GPT versions and Gemini 1.5 Flash displayed lower agreement (below 0.5) [41]. While LLMs perform well in rubric-based assessments, we noticed they struggle with discourse-level writing evaluation, particularly in areas requiring structural feedback, where human expertise is essential [34, 45].

Although the literature reveals promising results, we found that some challenges remain in writing classification due to LLMs prompt sensitivity and response variability. The literature posits that even slight variations in prompt phrasing can significantly impact classification

[2] https://ai.google.dev/palm_docs/palm

accuracy and consistency [6]. To address this, multi-step classification strategies have been proposed, where models first distinguish broad proficiency levels before refining subcategories [7, 22]. Bias in scoring also remains a concern. In an evaluation of 20 essays, for example, the GPT-4o1 model achieved the highest correlation with human ratings (Spearman's = .74) but, like GPT-4, exhibited a tendency to overrate texts [42].

LLM reliability varies with task complexity and domain. A study classifying 1,693 assessment questions based on Bloom's taxonomy [5] found traditional Machine Learning (ML) models outperforming Generative Artificial Intelligence (GenAI) models, with GPT-3.5-turbo achieving 0.62 precision versus 0.87 for traditional ML techniques, while Gemini Pro underperformed (0.43) [20]. Additionally, we learned that in [27] that Gemini Pro tends to struggle with fine-grained text comprehension tasks compared to GPT-4. Despite these limitations, LLMs have demonstrated potential in educational applications, particularly in tutoring. In an evaluation of 58,459 tutoring interactions, LearnLM[3] was preferred by expert raters over GPT-4o, Claude 3.5, and Gemini 1.5 Pro, highlighting its capacity for AI-driven learning support [45].

The studies we have here presented align with our objectives and research questions, emphasizing the complexities of assessing student writing consistently and accurately, and highlighting opportunities to improve traditional automated placement systems (e.g., ACCUPLACER) by leveraging LLMs and AI technologies. As these models evolve, improving classification precision, mitigating bias, and improving feedback quality will be paramount to ensure more reliable and equitable high-stakes assessments, particularly in DevEd [25, 28].

3 Methodology

To evaluate the precision of LLMs and their alignment with human raters for improving writing placement in DevEd, we first present a description of the corpus in Section 3.1, followed by a detailed explanation of the experimental setup in Section 3.2, along with an overview of the LLMs used and the classification experiments devised for this study.

3.1 Corpus

The corpus consists of essays written in English by college-intending students during the 2023–2024 academic year as part of Tulsa Community College's[4] standardized placement process. All essays were produced in a proctored environment without access to writing aids and were prompted by one of eleven possible reflective topics (e.g., *differences among people*, *unlimited change*) with minimal instructions.

A stratified random sampling method was employed to draw 450 essays from a larger pool of 1,000 placement essays. This sampling strategy minimized selection bias, complied with institutional data access protocols, and resulted in a corpus evenly distributed across three levels: DevEd Level 1, DevEd Level 2, and College-Level (150 texts each). The dataset combines an existing 300-essay developmental education corpus with 150 newly added College-Level texts, allowing for a more comprehensive evaluation of writing proficiency. Although demographic data was available for 67% of participants, it was excluded from the present analysis and will be examined in future research.

[3] https://ai.google.dev/gemini-api/docs/learnlm
[4] https://www.tulsacc.edu

For text-level classification, we reference an adapted version of the institution's official course descriptions:

- **DevEd Level 1**: Texts with frequent grammar, spelling, and punctuation issues, and lacking cohesion;
- **DevEd Level 2**: Texts with some structural improvements still needing targeted support; and
- **College Level**: Texts demonstrating academic-level writing with minimal errors.

All essays were also independently evaluated by two trained raters. Both raters are native English speakers who hold at least a bachelor's degree and have over five years of experience in higher education, particularly in DevEd curriculum and student writing assessment at U.S. community colleges. They were recruited through an open call for volunteers to participate in the writing assessment task. Prior to scoring, raters completed calibration sessions using classification guidelines developed by the authors of this paper [9] to ensure consistency in applying the three-level placement scale. Discrepancies were resolved through adjudication and score averaging, a widely accepted practice in classification tasks. Although only two raters were employed due to the intensive nature of the task and institutional resource constraints, their qualifications and training contributed to strong agreement. Interrater reliability [19] analysis yielded substantial agreement (K-alpha = 0.66), demonstrating a high level of consistency and reinforcing the dataset's reliability.

Table 1 presents key corpus statistics, including token distribution and classification assignments by Accuplacer and human raters.

Table 1 Corpus statistics and classification by level (by Accuplacer and human raters.

Levels	Tokens/text	Accuplacer	Human Raters
DevEd Level 1	≈ 190	150	118
DevEd Level 2	≈ 321	150	238
College Level	≈ 481	150	94
Total	**≈ 148,800**	**450**	**450**

It is pertinent to note that access to writing assessment corpora from standardized entrance exams is governed by strict protocols to protect at-risk, educationally disadvantaged student populations, ensure data privacy, and comply with institutional and ethical guidelines, which inherently limits their availability for research and broader analysis. This study adhered to these ethical considerations and followed the Institution's Review Board (IRB) protocols, under approval # 22-05.

3.2 Experimental Setup

We investigated the feasibility of using LLMs as complementary tools for placement decisions and writing instruction in DevEd. Specifically, we evaluated their classification performance against the gold standard corpus classification described in Section 3.1. Furthermore, the results of the LLM classification were then compared with the institution's existing placement system, Accuplacer, to assess its precision against alternative methods.

To evaluate classification performance, two classification experiments were envisaged:

- (i) a **1-step classification**, where students' texts were categorized into three classes: College-level, DevEd Level 1, and DevEd Level 2; and
- (ii) a **2-step classification**, where texts were first classified into College and DevEd categories and then further divided into DevEd Levels 1 and 2.

The transition from a **1-step** to a **2-step** classification focuses on potentially reducing initial complexity by first making a broad binary distinction (College vs. DevEd), ensuring that texts with clear college-level characteristics are correctly identified before refining classifications within the DevEd category. Since the differences between DevEd Levels 1 and 2 are often more subtle, this sequential approach allows for a more targeted classification between proficiency levels [18].

LLMs Selection

Four LLMs were tested and evaluated using their default hyperparameters. This study adopts the commonly accepted definition of large language models as pre-trained, transformer-based architectures [38]. Model selection was based on their wide availability via API access and their representation of varying tiers of generative performance, reasoning ability, and cost-efficiency – factors frequently highlighted in the literature as critical for educational applications and scalable implementation [18, 26, 47].

Model selection was also guided by practical considerations, including task complexity, processing speed, and data security – criteria commonly emphasized in applied LLM research and deployment frameworks. While the selected models align with current literature and offer broad applicability, this does not preclude the inclusion of additional or emerging models (e.g., DeepSeek) in future iterations of this research. The four LLMs included in this study are listed below in alphabetical order:

- **Claude**[5] [2], developed by Anthropic, is known to provide a more accessible API and structured responses, making it easier to evaluate for classification consistency [6]. It has also been reported to have better consistency in logical reasoning [4], which may enhance its performance in tasks requiring nuanced proficiency distinctions (e.g., DevEd Level 1 vs. Level 2). Claude 3.5 Sonnet (October 2024 version) was the model used.
- **Gemini**[6] [37], unlike the text-based LLMs GPT-3.5-turbo and GPT-4o, is designed for cross-modal reasoning and trained on diverse web sources, including educational content [44], which may enhance its ability to recognize academic proficiency markers and improve efficiency in handling texts with high linguistic variability [24, 27]. For this study, Gemini 2.0 Flash was used.
- **OpenAI's ChatGPT**[7]: (i) **GPT-3.5-turbo**, while computationally efficient [32], it is leveraged to assess any improvements in classification consistency with the more advanced, latest cost-efficient model [39], (ii) **GPT-4o**, in the task of DevEd placement context. The advanced model is explored for its strengths in text comprehension and structured response generation, which could not only help students identify writing weaknesses but also enhance the provision of targeted, context-aware feedback [27]. The experiments use the GPT models via the official OpenAI API.

LLMs Classification Experiments

To systematically examine how prompt structuring influences classification precision, we incorporated a four-stage prompting approach into the two experiments described in Section 3.2: (i) 1-step classification and (ii) 2-step classification.

[5] https://docs.anthropic.com/en/home
[6] https://deepmind.google/technologies/gemini/
[7] https://chat.openai.com

Prompt construction was guided by existing literature on LLM-based classification and rubric-based writing assessment [36, 48], ensuring that the design was not ad hoc. Prompts were iteratively refined to maximize linguistic clarity, align with defined classification objectives, and maintain compatibility across model architectures. These human-authored prompts were designed to capture patterns that distinguish proficient from non-proficient writing [31]. The complete series of prompts was archived for public access via GitHub[8] [11].

To evaluate whether increasing prompt specificity improved classification outcomes, we introduced a four-stage prompting sequence. Precision, the primary evaluation metric in this study, assessed the reliability of a model's positive predictions aligning with human-assigned classifications [21]. This focus is particularly relevant in DevEd contexts, where misclassification, whether through overplacement or underplacement, can significantly affect students' academic trajectories. The four prompting stages were as follows:

- (i) **Zero-shot**, in which the models classified texts following a prompt with a laconic approach. No sample texts were provided at this stage in the classification prompt. Due to the availability of comparable data, in terms of model precision, results from this approach establish the *baseline* for evaluation with the subsequent prompting strategies.
- (ii) **Few-shot**, in which the models received the same prompt but with basic definitions of each proficiency level, adapted from the Institution's official course descriptions and curriculum. No sample texts were provided at this stage either.
- (iii) **Enhanced**, in which the classification prompt included detailed descriptions of proficiency levels and key linguistic markers that encompassed grammatical accuracy, syntactic complexity, lexical richness, and discourse cohesion. No sample texts were provided at this stage.
- (iv) **Enhanced+**, this approach extends the **Enhanced** classification prompting by incorporating three manually assessed and classified sample texts (using ACCUPLACER's linguistic descriptors), each corresponding to the three proficiency levels already mentioned (DevEd Level 1, DevEd Level 2, and College-level). These samples were randomly selected from outside the corpus but drawn from the same database and time epoch.

The four-stage experimental framework aligned with the research questions in Section 1, allowing for a comparative analysis of how LLMs classify writing proficiency under varying levels of linguistic guidance. This approach evaluated classification performance with and without explicit linguistic criteria and concrete linguistic evidence (text samples). Across all prompting strategies, LLMs were also instructed to provide classification rationales, which, while more robust than ACCUPLACER, remain less comprehensive than human evaluations – an aspect to be explored in a subsequent study. A sample of this feedback is included in Appendix A.

4 Experimental Results

This section presents the results of the two classification experiments[9] described in Section 3.2: (i) 1-step classification and (ii) 2-step classification. In Sections 4.1 and 4.2, we analyze LLMs' performance across the four-stage prompting approaches, evaluating their effectiveness in distinguishing proficiency levels.

[8] `https://gitlab.hlt.inesc-id.pt/u000803/deved`
[9] All experiments were conducted over eight days, from February 18 to February 25, 2025.

4.1 1-step Classification: 3 classes

The 1-step classification experiment required LLMs to simultaneously differentiate among three proficiency levels – College-Level, DevEd Level 1, and DevEd Level 2 – without an intermediate binary decision. This method directly assessed the models' ability to distinguish college-ready writing from varying degrees of developmental writing in a single step.

Table 2 presents each LLMs' precision in classifying texts across these three levels under four prompting strategies. The results include precision figures for each level compared to the gold standard, as well as the overall model performance. A "+" next to the overall precision score indicates improvement relative to the Zero-shot approach (baseline), while a "−" denotes a decline. Results are analyzed in terms of overall precision first before comparing them to ACCUPLACER. Additionally, a confusion matrix of the best-performing model is included to provide a more detailed breakdown of classification outcomes relative to the gold standard.

Table 2 LLM and ACCUPLACER classification precision vs. gold standard, per level and overall.

LLM / Experiment	Level 1	Level 2	College	Overall Precision
Claude Zero-shot	44.44%	60.00%	72.09%	52.44%
Claude Few-shot	53.65%	61.73%	86.67%	59.1% (+6.66%)
Claude Enhanced	50.00%	66.83%	85.71%	60.89% (+8.45%)
Claude Enhanced+	50.89%	73.65%	73.08%	**62.22** % (**+9.78**%)
Gemini Zero-shot	41.16%	70.00%	64.38%	51.33%
Gemini Few-shot	42.86%	60.12%	81.08%	52.44% (+1.11%)
Gemini Enhanced	29.29%	64.04%	84.62%	42.89% (-8.44%)
Gemini Enhanced+	52.53%	63.76%	78.26%	59.56% (+8.23%)
GPT-3.5-turbo Zero-shot	28.78%	20.69%	81.82%	29.56%
GPT-3.5-turbo Few-shot	38.75%	65.04%	68.42%	48.44% (+18.88%)
GPT-3.5-turbo Enhanced	55.33%	59.26%	40.54%	53.33% (+23.77%)
GPT-3.5-turbo Enhanced+	43.85%	57.09%	43.75%	51.11% (+21.55%)
GPT-4o Zero-shot	36.59%	49.14%	88.24%	41.78%
GPT-4o Few-shot	43.95%	65.38%	78.26%	54.89% (+13.11%)
GPT-4o Enhanced	41.54%	69.53%	74.19%	54.00% (+12.22%)
GPT-4o Enhanced+	49.09%	70.76%	79.66%	61.33% (+19.55%)
Accuplacer	51.33%	68.67%	54.67%	58.22%

In the **Zero-shot** experiment, Claude achieved the highest classification precision (52.44%), followed closely by Gemini (51.33%), outperforming all other LLMs. The remaining models ranked in descending order of precision as GPT-4o > GPT-3.5-turbo.

In the **Few-shot** experiment, when given basic proficiency definitions, Claude remained the best-performing model (59.1%), showing noticeable gains (approximately +6.66%) from the Zero-shot run. The ranking of models in this experiment was: GPT-4o > Gemini > GPT-3.5-turbo.

In the **Enhanced** experiment, when refined linguistic descriptors were introduced, Claude maintained its leading position (60.89%). However, not much improvement (+1.79%) was observed from the previous setup (Claude Few-shot). The remaining models ranked as follows: GPT-4o > GPT-3.5-turbo > Gemini.

Lastly, in the **Enhanced+** experiment, Claude continued to perform best (62.22%), with a 10.22% improvement over its Zero-shot outcome, but only +1.33% improvement over the previous setup. The ranking, this time, was GPT-4o > Gemini > GPT-3.5-turbo. Notably, the highest gains in precision from the Zero-shot to the enhanced+ experiments were seen in GPT-3.5-turbo (21.55%) and GPT-4o (19.55%).

When compared to ACCUPLACER, three models, Claude (4%), GPT-4o (3.11%), and Gemini (1.34%), achieved slightly higher overall precision in the Enhanced+ experiments. Notably, Claude outperformed ACCUPLACER also across the Few-shot and Enhanced prompting strategies. While LLMs show potential for structured DevEd placement tasks compared to ACCUPLACER, they still present limitations, as evidenced by their low precision scores.

To complement this analysis, Figure 1 provides a visual representation of the classification precision scores distribution of LLMs and ACCUPLACER in the 1-step classification experiment.

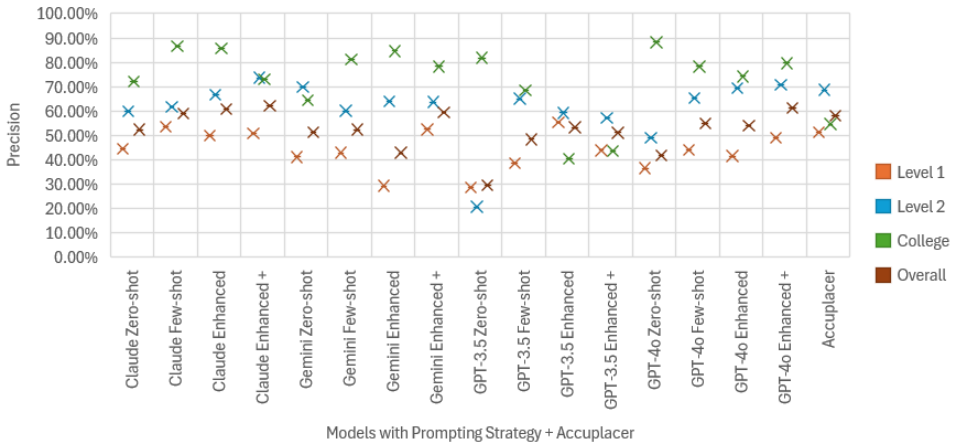


Figure 1 LLM and ACCUPLACER classification precision distribution per level and overall.

To better understand Claude's overall performance, given its highest classification precision in all prompting strategies, Table 3 displays the confusion matrix with the model's distribution of actual versus predicted classifications in its best performance, the enhanced+ experiment:

Table 3 Confusion matrix of predicted (LLM) vs. actual classifications (gold standard) of Claude in the Enhanced+ experiment.

↓ Actual / Predicted →	Level 1	Level 2	College	Total
Level 1	**114**	4	0	118
Level 2	108	**109**	21	238
College	2	35	**57**	94
Total	224	148	78	450

The confusion matrix analysis revealed that Claude excels in identifying Level 1 texts, correctly classifying nearly all (114/118) with only four misclassified as Level 2, indicating a strong grasp of beginning-level writing. However, its performance dropped significantly for Level 2, correctly identifying less than half of these texts. A strong tendency to underclassify was observed, as 108 Level 2 texts are misclassified as Level 1, and 21 as College-Level. Claude demonstrated good discrimination of advanced writing, correctly classifying 57 out of 94 College-Level texts, with most misclassifications occurring just one level below. Extreme errors were rare, with only two College-Level texts misclassified as Level 1. Overall, the model underclassified far more often than it overclassified (145 vs. 25 instances), suggesting it applied a stricter threshold for advancement compared to the human gold standard.

Based on the classification results obtained, further statistical validation is necessary. Pearson correlation coefficients (ρ) are computed next to assess the reliability and alignment of LLM-generated classifications with the gold standard.

Correlation of LLMs Classifications with Gold Standard

Pearson correlation coefficients (ρ) measured the degree of linear association between the classification levels assigned by LLMs and the gold standard in the 1-step experiment. A higher ρ value indicates a stronger correlation, meaning the LLMs' predictions align more closely with human-assigned levels.

Pearson correlation (ρ) scores were interpreted using SPSS guidelines[10], based on Cohen (1988) [8], which define correlation strength as follows:

$0.1 < |r| < 0.3$ indicates *small/weak* correlation (W);
$0.3 < |r| < 0.5$ corresponds to *medium/moderate* correlation (M); and
$0.5 < |r| \ldots$ denotes *large/strong* correlation (S).

The computed Pearson correlation coefficients (ρ) for each system are presented in Table 4.

Table 4 Pearson scores of LLMs and Accuplacer vs. gold standard. Correlation interpretation: substantial (S); moderate (M).

LLM/System	Zero-shot	Few-shot	Enhanced	Enhanced+
Claude	0.67 (S)	0.66 (S)	0.67 (S)	**0.75 (S)**
Gemini	0.60 (S)	0.62 (S)	0.62 (S)	0.63 (S)
GPT-3.5-turbo	0.40 (M)	0.54 (S)	0.51 (S)	0.39 (M)
GPT-4o	0.63 (S)	0.60 (S)	0.60 (S)	0.69 (S)
Accuplacer	**0.67 (S)**			

The results indicated that most LLMs exhibited a strong correlation with the gold standard, with Claude Enhanced+ achieving the highest alignment ($\rho = 0.75$), surpassing Accuplacer ($\rho = 0.67$). GPT-4o also outperformed Accuplacer but just by a smaller margin (0.02%) in the Enhanced+ run as well. In the Zero-shot and Enhanced prompting strategies, Claude matched Accuplacer's Pearson score, while the remaining models and configurations consistently underperformed.

Notably, GPT-3.5-turbo Zero-shot showed one of the weakest correlation scores ($\rho = 0.40$), but its performance improved significantly in the Few-shot ($\rho = 0.54$) and Enhanced ($\rho = 0.51$) settings before declining again in Enhanced+ ($\rho = 0.39$). This fluctuation suggests that models such as GPT-3.5-turbo may be more sensitive to prompt variations than others.

The strong correlation observed with Claude Enhanced+, particularly in such structured classification settings, highlights the potential of certain LLMs to approximate human-level proficiency in classification, though inconsistencies across models require further investigation.

4.2 2-step Classification: 3 classes

The 2-step classification experiment is followed in an attempt to address the risk of models disproportionately assigning texts to a particular category. Dividing a classification problem into two steps has shown to improve the overall precision of certain LLMs [29, 48]. To this extent, all 450 texts were initially classified as either College-Level or DevEd Level. Then, the DevEd-level texts, specifically, underwent a second classification, further distinguishing between DevEd Level 1 and DevEd Level 2. While the prompting methodology remained consistent (compared to the 1-step), level definitions and provided examples were slightly adjusted to align with the objectives of this classification experiment.

[10] `https://libguides.library.kent.edu/SPSS/PearsonCorr`

Table 5 presents the final precision scores[11] for each LLM (per level) across all four prompting strategies. A "+" next to the overall precision score indicates improvements relative to the Zero-shot approach (baseline), while a "−" represents the opposite. [13]

Table 5 LLM and ACCUPLACER classification precision vs. gold standard, per level and overall.

LLM / Experiment	Level 1	Level 2	College	Overall Precision
Claude Zero-shot	46.19%	74.07%	63.21%	56.89%
Claude Few-shot	51.72%	76.67%	57.48%	60.00% (+3.11%)
Claude Enhanced	52.88%	77.60%	58.97%	61.33% (+4.44%)
Claude Enhanced+	57.95%	77.36%	62.61%	66.00 % (+9.11%)
Gemini Zero-shot	48.83%	60.37%	82.61%	56.07%
Gemini Few-shot	57.79%	63.33%	76.92%	62.22% (+6.15%)
Gemini Enhanced	54.49%	61.92%	73.91%	59.78% (+3.71%)
Gemini Enhanced+	**66.67%**	**69.50%**	**73.85%**	**69.33% (+13.26%)**
GPT-3.5-turbo Zero-shot	45.29%	58.97%	43.62%	47.11%
GPT-3.5-turbo Few-shot	48.80%	64.52%	54.22%	50.89% (+3.78%)
GPT-3.5-turbo Enhanced	37.20%	65.00%	61.40%	46.44% (-0.67%)
GPT-3.5-turbo Enhanced+	46.24%	62.76%	47.06%	53.56% (+6.44%)
GPT-4o Zero-shot	40.81%	59.26%	72.09%	49.33%
GPT-4o Few-shot	34.24%	48.54%	82.35%	39.33% (-10.00%)
GPT-4o Enhanced	35.09%	51.38%	89.47%	41.33% (-8.00%)
GPT-4o Enhanced+	48.91%	73.83%	73.61%	61.11% (+11.78%)
Accuplacer	51.33%	68.67%	54.67%	58.22%

In the **Zero-shot** experiment, Claude demonstrated the highest classification precision (56.89%), closely followed by Gemini with only a 0.82% difference. The ranking of the remaining models in decreasing precision is GPT-4o > GPT-3.5-turbo.

In the **Few-shot** experiment, Gemini outperformed Claude by 2.22%. Both models showed modest improvements over their Zero-shot runs, with Claude achieving a 3.11% increase and Gemini, with almost twice the precision gain, achieving a 6.15%. The ranking of the remaining models was reversed from the Zero-shot experiment, as GPT-3.5-turbo outperformed GPT-4o. Notably, GPT-4o exhibited a significant 10% decrease in precision, indicating that added proficiency definitions did not necessarily enhance its performance.

In the **Enhanced** experiment, Claude achieved the highest precision score (61.33%), outperforming all other models. However, its improvement over the Few-shot setup was minimal (+1.33%), despite the inclusion of more precise linguistic descriptors and features in the prompt. Gemini ranked second, showing a 3.71% gain from its baseline but a 2.44% drop from the Few-shot setup, suggesting inconsistent benefits from increased prompt specificity. The remaining models ranked as follows: GPT-3.5-turbo > GPT-4o.

In the **Enhanced+** experiment, Gemini demonstrated the most significant gains, achieving an overall precision of 69.33%. Claude followed closely with a precision score of 66%, while GPT-4o ranked third with a precision score of 61.11%. GPT-3.5-turbo had the lowest precision in this setup, scoring 53.56%. All four models demonstrated performance gains from each of their baselines (Zero-shot) and the Enhanced run. However, Claude was the model to consistently improve its precision across all prompting strategies, suggesting that incorporating text samples alongside linguistic guidance enhances classification precision.

[11] This study is part of an ongoing research project. While the textual portion of the corpus cannot be released at this stage, a dataset containing the full classification results from the 2-step experiment – where overall precision scores exceeded those of the 1-step approach – has been made available for evaluation at `https://gitlab.hlt.inesc-id.pt/u000803/deved/` [13]. Full corpus access will be available upon completion of the study.

When compared to ACCUPLACER, Gemini and Claude demonstrated the highest overall precision in the Enhanced+ experiments, with notable gains of 11.11% and 7.78%, respectively. Both models also outperformed ACCUPLACER in the Few-shot and Enhanced strategies. GPT-4o showed a smaller improvement of 2.89% in the Enhanced+ strategy, while GPT-3.5-turbo consistently underperformed throughout. Despite their enhanced potential for structured DevEd placement, LLMs still exhibit some limitations, as reflected in their precision scores.

To complement this analysis, Figure 2 provides a visual representation of the classification precision scores distribution of LLMs and ACCUPLACER in the 2-step classification approach.

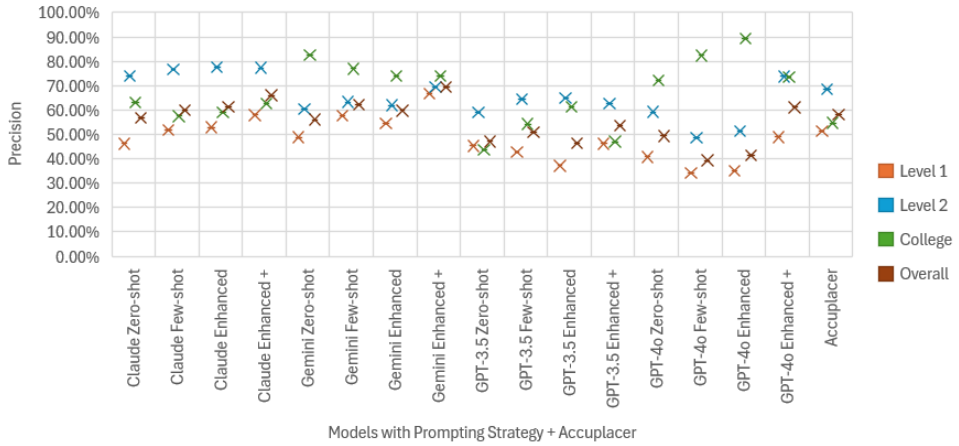


Figure 2 LLM and ACCUPLACER classification precision distribution, per level and overall.

Table 6 shows a confusion matrix detailing the model's distribution of actual versus predicted classifications in its best performance, the Enhaced+ experiment:

Table 6 Confusion matrix of predicted (LLM) vs. actual classifications (gold standard) in the Gemini Enhanced+ experiment.

↓ Actual / Predicted →	Level 1	Level 2	College	Total
Level 1	**84**	33	1*	118
Level 2	42	**180**	16	238
College	0	46	**48**	94
Total	126	259	65	450

Almost all errors (137 out of 138 misclassifications) occurred between adjacent levels, suggesting the model understands the ordinal relationship between levels. When Gemini made errors on Level 1 texts, it nearly always overclassified texts by just one level. The single outlier*, a text classified by Gemini as College-level but by human raters as DevEd Level 1, was closely examined. This text contained a single string of tokens, yet it was the shortest in the dataset.

Text 13: "Our ability to change ourselves is limitless."

Notably, Gemini distinguished it as a "statement" rather than a "text," which is the terminology used in all other justifications. The model's justification *verbatim* was:

> This is a concise, grammatically correct *statement* expressing a complex idea. It demonstrates strong control of language and does not contain any foundational errors. The sentence structure is simple but effective, conveying a clear and impactful message.

Gemini, additionally, showed some bias toward classifying texts as Level 2, which is also the level involved in most errors, both as the actual and predicted class. Overall, Gemini is more likely to underclassify than overclassify Level 2 texts, suggesting Level 2 is the most difficult to distinguish. On the contrary, the model never misclassified College texts as Level 1, showing excellent discrimination between these distant categories. As exhibited in Table 5, in the Enhanced+ run, Gemini maintains similar precision across all levels (≈67-74%), indicating balanced performance rather than excelling at one level at the expense of others.

Correlation of LLMs Classifications with Gold Standard

Similarly to the 1-step experiment, Pearson correlation coefficients (ρ) were also computed. Results are presented in Table 7.

Table 7 Pearson scores of LLMs and ACCUPLACER vs. gold standard. Correlation interpretation: substantial (S); moderate (M).

LLM/System	Zero-shot	Few-shot	Enhanced	Enhanced+
Claude	0.68 (S)	0.69 (S)	0.71 (S)	**0.73 (S)**
Gemini	0.64 (S)	0.62 (S)	0.58 (S)	0.69 (S)
GPT-3.5-turbo	0.59 (M)	0.54 (S)	0.49 (M)	0.45 (M)
GPT-4o	0.61 (S)	0.49 (M)	0.54 (S)	0.70 (S)
Accuplacer	**0.67 (S)**			

Compared to the 1-step classification, Pearson correlation calculations revealed that most LLMs exhibit a strong alignment with the gold standard. Claude Enhanced+ achieved the highest correlation ($\rho = 0.73$), followed closely by Claude Enhanced ($\rho = 0.71$) and GPT-4o ($\rho = 0.70$), outperforming ACCUPLACER ($\rho = 0.67$). There was a minimal difference (0.02% points) in Pearson scores observed between the 1-step and 2-step classifications for Claude Enhanced+, hinting at the model's stability in its classification performance.

The strong correlation observed in Claude and GPT-4o reinforces the idea that LLMs, especially in a 2-step classification, can closely align with human-assigned proficiency levels.

4.3 Precision Comparison: 1-step vs. 2-step classification experiments

Having conducted both 1-step and 2-step classification experiments using four prompting strategies, we highlight in Table 8 the top 3 major precision gains (green) and losses (red), per level and overall, to assess how LLMs respond to structured classification strategies. The figures represent the difference in precision between the 2-step and 1-step experiments, as overall precision scores were higher in the 2-step. This comparison directly supports the analysis of research questions 2 and 3 in Section 1, evaluating the impact of prompt refinement on classification precision and bias mitigation.

Most models improved in overall precision with the 2-step classification, particularly GPT-3.5-turbo Zero-shot (+17.55%), Gemini Enhanced (+16.89%), and Gemini Few-shot (+9.78%). Although not among the top three, Gemini Enhanced+ performed similarly to Gemini Few-shot, with only a 0.01% difference. Conversely, GPT-4o struggled to maintain positive precision gains across three out of the four prompting strategies, suggesting that a 2-step classification strategy may not be as effective for this particular model.

We observed significant gains in Level 1 in two of the four prompting strategies, particularly for Gemini (Few-shot: +14.93%, Enhanced: +25.20%) and GPT-3.5-turbo Zero-shot (+16.51%). Particularly for GPT-3.5-turbo Zero-shot, it also saw the sharpest trade-off,

Table 8 Precision gains and losses between 1 and 2-step experiments across prompting strategies.

LLM / Experiment	Level 1	Level 2	College	Overall
Claude Zero-shot	+1.75%	+14.07%	-8.88%	+4.45%
Claude Few-shot	-1.93%	+14.94%	-29.19%	+0.90%
Claude Enhanced	+2.88%	+10.77%	-26.74%	+0.44%
Claude Enhanced+	+7.06%	+3.71%	-10.47%	+3.78%
Gemini Zero-shot	+7.67%	-9.63%	+18.23%	+4.74%
Gemini Few-shot	+14.93%	+3.21%	-4.16%	+9.78%
Gemini Enhanced	+25.20%	-2.12%	-10.71%	+16.89%
Gemini Enhanced+	+14.14%	+5.74%	-4.41%	+9.77%
GPT-3.5-turbo Zero-shot	+16.51%	+38.28%	-38.20%	+17.55%
GPT-3.5-turbo Few-shot	+4.05%	-0.52%	-14.20%	+2.45%
GPT-3.5-turbo Enhanced	-18.13%	+5.74%	+20.86%	-6.89%
GPT-3.5-turbo Enhanced+	+2.39%	+5.67%	+3.31%	+2.45%
GPT-4o Zero-shot	+4.22%	+10.12%	-16.15%	+7.55%
GPT-4o Few-shot	-9.71%	-16.84%	+4.09%	-15.56%
GPT-4o Enhanced	-6.45%	-18.15%	+15.28%	-12.67%
GPT-4o Enhanced+	-0.18%	+3.07%	-6.05%	-0.22%

improving +38.28% in Level 2 but dropping nearly the same amount (-38.20%) in College-Level classification. Similarly, for Level 2, Claude Zero and Few-shot showed notable gains (+14.07% and +14.94%, respectively) but substantial losses at the College-Level (-29.19% and -26.74%) in the Few-shot and Enhanced runs.

Overall, most results suggest that breaking down classification into distinct steps improved precision. Among all models, Gemini demonstrated the most consistent performance across levels and prompting strategies. The trade-off between gains in lower proficiency levels and losses in College-Level precision suggests that some models, particularly Claude and GPT-3.5-turbo, may be more in-tuned with DevEd writing patterns, prioritizing features typical of lower-level texts while misclassifying more advanced writing as DevEd.

5 Conclusions and Future Work

We evaluated four LLMs as data-driven alternatives for equitable DevEd placement, offering (potential) competitive, cost-effective suggestions for higher education institutions traditionally reliant on "black-box" systems like ACCUPLACER. Our findings indicated that both classification structure and prompt refinement played a critical role in enhancing LLM precision when distinguishing between DevEd and College-level texts.

In terms of classification structure, across the 1 and 2-steps classification experiments, models exhibited higher precision in Level 2 and College-Level classifications but struggled with Level 1, likely due to the inconsistent structures and frequent grammatical errors in Level 1 texts. Compared to the 1-step experiment, the 2-step classification saw Gemini Enhanced+ achieved the highest precision (66.67%) in Level 1, reinforcing the notion that breaking classification into multiple steps enhances precision [29, 48], and the importance of incorporating clear linguistic criteria with sample texts in the classification prompts [36]. We noticed, too, that Claude and Gemini consistently underclassified Level 2 texts, suggesting they apply a stricter advancement threshold compared to human raters. This likely occurs because the models are more conservative in assigning texts to higher proficiency levels, meaning they are more prone to classifying Level 2 texts as Level 1 rather than correctly identifying them.

Claude and Gemini demonstrated a strong correlation with human ratings, with Claude achieving the highest Pearson scores ($\rho = 0.75$; 1-step, $\rho = 0.73$; 2-step), surpassing ACCUPLACER ($\rho = 0.67$). These findings confirm that LLMs have the potential approximate human-level proficiency classification, yet their precision remains a work in progress, particularly for high-stakes placement decisions. Given the social and economic impact of misplacement, safeguards are necessary to ensure an ethical and fair assessment of students' skills.

To refine the role of LLMs in placement decisions, our future research will: (i) explore the expansion of the dataset to increase generalizability; (ii) further refine prompting strategies to mitigate classification inconsistencies; (iii) test newer, more sophisticated LLMs for improved performance; (iv) investigate LLM-generated feedback to evaluate its effectiveness in justifying classification decisions and enhancing instructional usability in teaching practices. By addressing these areas, future studies will continue to contribute to more transparent, data-driven placement methodologies that balance automation with fairness and reliability.

References

1 Tufan Adiguzel, Mehmet Haldun Kaya, and Faith Kürşat Cansu. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3):ep429, 2023. doi:10.30935/cedtech/13152.

2 AI Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. *Claude-3 Model Card*, 1:1, 2024.

3 Nathan Atox and Mason Clark. Evaluating Large Language Models through the lens of linguistic proficiency and world knowledge: A comparative study. *Authorea Preprints*, 2024.

4 Jaehyeon Bae, Seoryeong Kwon, and Seunghwan Myeong. Enhancing software code vulnerability detection using gpt-4o and claude-3.5 sonnet: A study on prompt engineering techniques. *Electronics*, 13(13), 2024. doi:10.3390/electronics13132657.

5 Benjamin Samuel Bloom. *Taxonomy of educational objectives: Handbook II*. David McKay, 1956.

6 Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. Claude 2.0 Large Language Model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications*, 21:200336, 2024. doi:10.1016/J.ISWA.2024.200336.

7 Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2778–2788, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3485447.3511998.

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

9 Miguel Da Corte and Jorge Baptista. Enhancing writing proficiency classification in Developmental Education: the quest for accuracy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, 2024. URL: https://aclanthology.org/2024.lrec-main.542.

10 Miguel Da Corte and Jorge Baptista. Leveraging NLP and machine learning for English (l1) writing assessment in Developmental Education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140, 2024. doi:10.5220/0012740500003693.

11 Miguel Da Corte and Jorge Baptista. LLM-based writing classification prompts for 1-step and 2-step classification experiments. Technical report, University of Algarve & INESC-ID Lisboa, 2025.

12 Miguel Da Corte and Jorge Baptista. Refining English writing proficiency assessment and placement in Developmental Education using NLP tools and Machine Learning. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 288–303. INSTICC, SciTePress, 2025. doi:10.5220/0013351500003932.

13 Miguel Da Corte and Jorge Baptista. Results for LLM-based 2-step classification. Technical report, University of Algarve & INESC-ID Lisboa, 2025.

14 Miguel Da Corte and Jorge Baptista. Toward consistency in writing proficiency assessment: Mitigating classification variability in Developmental Education. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 139–150. INSTICC, SciTePress, 2025. doi:10.5220/0013353900003932.

15 Jane Denison-Furness, Stacey Lee Donohue, Annemarie Hamlin, and Tony Russell. 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 Postsecondary Education*, pages 107–127. The WAC Clearinghouse/University Press of Colorado, 2022. doi:10.37514/PRA-B.2022.1565.2.04.

16 Nikki Edgecombe and Michael Weiss. Promoting equity in Developmental Education reform: A conversation with Nikki Edgecombe and Michael Weiss. *Center for the Analysis of Postsecondary Readiness*, page 1, 2024.

17 Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. GPT-3.5, GPT-4, or BARD? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032, 2023. doi:10.1016/J.NLP.2023.100032.

18 John Fields, Kevin Chovanec, and Praveen Madiraju. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531, 2024. doi:10.1109/ACCESS.2024.3349952.

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

20 Mohammed Osman Gani, Ramesh Kumar Ayyasamy, Saadat M Alhashmi, Khondaker Sajid Alam, Anbuselvan Sangodiah, Khondaker Khaleduzzman, and Chinnasamy Ponnusamy. Towards enhanced assessment question classification: a study using Machine Learning, deep learning, and Generative AI. *Connection Science*, 37(1):2445249, 2025. doi:10.1080/09540091.2024.2445249.

21 Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020. arXiv:2008.05756.

22 Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with Large Language Models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023. URL: https://proceedings.mlr.press/v206/hegselmann23a.html.

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

24 Muhammad Imran and Norah Almusharraf. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22, 2024. doi:10.1186/S40561-024-00310-Z.

25 Elizabeth Kopko, Jessica Brathwaite, and Julia Raufman. The next phase of placement reform: Moving toward equity-centered practice. research brief. *Center for the Analysis of Postsecondary Readiness*, 2022. URL: https://files.eric.ed.gov/fulltext/ED626145.pdf.

26 Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. LLMs in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147, 2024. doi:10.1609/AAAISS.V3I1.31193.

27 Gyeong-Geon Lee, Ehsan Latif, Lehong Shi, and Xiaoming Zhai. Gemini Pro defeated by GPT-4v: Evidence from education. *arXiv preprint arXiv:2401.08660*, 2023.

28 Stephanie Link and Svetlana Koltovskaia. *Automated Scoring of Writing*, pages 333–345. Springer International Publishing, Cham, 2023. doi:10.1007/978-3-031-36033-6_21.

29 Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35, 2023. doi:10.1145/3560815.

30 Ross Markle. Redesigning course placement in service of guided pathways. *Educational Considerations*, 50(2):8, 2025.

31 Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in Natural Language processing via Large pre-trained Language Models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023. doi:10.1145/3605943.

32 Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia, May 2024. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.1055/.

33 Sasha Nikolic, Carolyn Sandison, Rezwanul Haque, Scott Daniel, Sarah Grundy, Marina Belkina, Sarah Lyden, Ghulam M Hassan, and Peter Neal. ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 29(2):126–153, 2024.

34 Austin Pack, Alex Barrett, and Juan Escalante. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234, 2024. doi:10.1016/J.CAEAI.2024.100234.

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

36 Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker, editors, *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia, May 2024. ELRA and ICCL. URL: https://aclanthology.org/2024.mwe-1.22/.

37 Sundai Pichai and Demis Hassabis. Introducing Gemini: our largest and most capable AI model, 2023. Accessed: February 19, 2025. URL: https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini.

38 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

39 Areeg Fahad Rasheed, M Zarkoosh, Safa F Abbas, and Sana Sabah Al-Azzawi. TaskComplexity: A dataset for task complexity classification with in-context learning, FLAN-T5 and GPT-4o benchmarks. *arXiv preprint arXiv:2409.20189*, 2024. doi:10.48550/arXiv.2409.20189.

40 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17, 2024. URL: https://aclanthology.org/2024.propor-2.4/.

41 Siti Bealinda Qinthara Rony, Tan Xin Fei, and Sasa Arsovski. Educational justice. reliability and consistency of Large Language Models for automated essay scoring and its implications. *Journal of Applied Learning and Teaching*, 8(1), 2025.

42 Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472, 2025. doi:10.1145/3706468.3706527.

43 Arthur Spirling. Why open-source Generative AI models are an ethical way forward for science. *Nature*, 616(7957):413–413, 2023.

44 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

45 LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

46 Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. Zero-shot text classification with knowledge resources under label-fully-unseen setting. *Neurocomputing*, 610:128580, 2024. doi:10.1016/j.neucom.2024.128580.

47 Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. Adaptable and reliable text classification using Large Language Models. *arXiv preprint arXiv:2405.10523*, 2024.

48 Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. Human-AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 293–305, 2025. doi:10.1145/3706468.3706507.

A Accuplacer and LLMs' sample feedback

Test Results

Test Name	Score
[redacted]	2

Holistic Score Description	
The essay demonstrates very little mastery of on-demand essay writing.	
Dimension Title	**Dimension Description**
Purpose and Focus	Your response lacks clear purpose and focus.
Organization and Structure	Your response demonstrates limited organization of ideas.
Development and Support	Your response has limited support for your ideas.
Sentence Variety and Style	Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Mechanical Conventions	Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking	Your response shows insufficient reasoning and lacks complexity of thought.

Course Placements - Determined on 07-20-2023

Writing Foundations I

Figure 3 Accuplacer's feedback for DevEd Level 1 = Writing Foundations I.

Gemini Enhanced+'s Feedback. Text: A 322; Classification: DevEd Level 1 Rationale:

> The text exhibits frequent errors in mechanical conventions, including misspellings ("themselfs," "simpley," "mojority," "to," "garentee," "sucurty," "revanew," "Anthoner," "belive," "unles"), punctuation (incorrect comma usage, missing apostrophes), and grammatical errors (subject-verb agreement, pronoun usage). Sentence variety and style are limited, with repetitive sentence structures. Development and support are present but lack coherence and depth. Organization and structure are attempted but weakened by abrupt transitions and lack of a clear thesis in the introduction. While the purpose and focus are identifiable, the arguments lack sufficient clarity. The text exhibits limited critical thinking, relying heavily on anecdotal evidence and generalizations.

This page was left blank intentionally.

Conclusion

This study was conducted during a period of significant change in higher education, first shaped by the aftermath of the COVID-19 pandemic and later by the emergence of newer Large Language Models (LLMs) such as ChatGPT (Chen et al. (2020); Wang et al. (2024a); McDonald et al. (2025)). These developments have since then prompted institutions to rethink learning environments and assessment practices (Engsig et al. (2025); Nieminen (2025)), creating both challenges and opportunities for innovation. Amid these changes, this thesis set out to answer a question pertaining to the fields of linguistics and education: *How can the writing of community college students be more accurately and equitably assessed, based on their written productions, to determine whether they require remediation or are ready for college-level coursework*? This question was addressed through corpus development and feature-based analysis, with the goal of improving placement decisions and supporting fairer educational outcomes.

Guided by this question, this thesis investigated how linguistically grounded features, both human-identified and automatically extracted, could improve the accuracy, interpretability, and fairness of writing assessment and placement in a U.S. community college. By analyzing the lexical and syntactic patterns of English (L1) speakers placed into DevEd courses, this research sought to: (i) build a clearer linguistic profile of developmental-level writing; (ii) evaluate how these features vary across proficiency levels; and (iii) assess their predictive value in classification tasks. Results from this thesis work responded to the pressing need for placement systems to provide a more transparent, linguistically informed assessment, compared to current automated tools, grounded in how students actually write. The resulting studies and contributions were presented in four thematic sections throughout this document, comprising nine double-blind, peer-reviewed papers.

At the institution where this study took place, Tulsa Community College (TCC), ACCUPLACER serves as the primary placement tool, classifying students based on short essay responses to various prompts. Beyond TCC, ACCUPLACER is also widely used across community colleges in the U.S. One of its main limitations, as mentioned throughout this thesis, is that it operates as a "black-box" algorithm, meaning its placement logic and the features considered in the assessment and placement decision are not publicly disclosed. This lack of linguistic transparency carries several pedagogical implications, as it limits the ability of faculty and institutional support services to target specific areas of writing where students may need assistance. In response, several contributions of this thesis directly address the concerns regarding the validity and accuracy of ACCUPLACER's results, as well as the broader implications these have for higher education, particularly within community colleges (Kopko et al. (2022); Kubiak (2022)).

English (L1) writing served as the evidentiary basis of this thesis, providing a comprehensive indicator of academic readiness within the context of higher education in the U.S. The ability to structure ideas coherently, use language precisely, and engage in meaningful argumentation is foundational for college success. Accordingly, this thesis addressed four interrelated challenges:

(i) the limited representation of DevEd writing in established corpora; (ii) the lack of transparency and linguistic grounding in ACCUPLACER's placement process; (iii) the high rate of misplacement by this system; and (iv) the absence of actionable linguistic feedback produced by ACCUPLACER upon assessment of a students' essays. These challenges are addressed in alignment with the research questions posed in this study.

To address the first challenge, (Da Corte and Baptista (2024a)), guided by a Corpus Linguistics approach (McEnery et al. (2006); Xiao (2010); Durrant (2022)), assembled a 100-essay corpus of student writing that had been initially automatically classified by ACCUPLACER. The texts were randomly selected from TCC's exam database and then securely stored in a dedicated research database created for this entire investigation. The corpus was balanced by placement levels (DevEd Level 1 and DevEd Level 2), which, at the time this study was conducted, were the two tiers of developmental writing courses offered at said institution. This corpus served as the foundation for a detailed linguistic annotation, with the specific aim of identifying linguistic markers characteristic of DevEd student writing.

Following a set of guidelines crafted by two expert linguists, a structured annotation protocol was developed. Prior to deploying the full annotation task, two expert linguists conducted a pilot study using three sample essays drawn from the same exam database and the same epoch. This pilot[54] was designed to identify and refine the linguistic features that would guide the annotation task. A total of 21 features were identified across the three sample essays and systematically defined, anchored in the literature and in the specific needs of the DevEd context. These features spanned orthographic, grammatical, lexical, and discursive patterns. Some of them highlighted areas of writing in need of further development, while others signaled more advanced (proficient) writing markers. These features were incorporated into a set of annotation guidelines, after which the 100 essays, comprising the initial corpus, were manually annotated. This process was carried out through a carefully designed annotation task by trained human raters who underwent extensive preparation and adhered to TCC's Institutional Review Board (IRB) protocols.

Before shifting focus to the language assessment aspect of this thesis, particularly as it relates to ACCUPLACER's placement process, it was necessary first to evaluate the linguistic framework and annotation protocols developed as an initial step in this study. This evaluation, which constitutes a key methodological contribution of the thesis, focused on calculating inter-rater agreement metrics to assess whether the devised linguistic features and descriptors were consistently applied or required further refinement.

As previously noted, extensive calibration work was undertaken, including the systematic selection and vetting of annotators. Six raters were initially recruited through an open call for volunteers and grouped into three pairs. Through a series of experiments reported in (Da Corte

[54]Preliminary findings from this pilot study were presented at the 14th International Conference on Corpus Linguistics (CILC 2023), held in May 2023 at the University of Oviedo. The abstract submitted is available in the conference's book of abstracts.

and Baptista (2024b)), the best-performing pair was identified using Intraclass Correlation Coefficients (ICCs) (Koo and Li (2016)). The two annotators in the best-performing pair, one male and one female, both worked in academic advising at TCC, held at least a Bachelor's degree, and were familiar with institutional placement guidelines and the distinctions between DevEd levels. They supported all subsequent experiments related to the annotation and classification tasks in this investigation. Their training took place over multiple sessions: the first session introduced the guidelines, the second involved supervised practice with training texts, and a final post-annotation session addressed any remaining ambiguities and reinforced consistency. Following this extensive preparation, inter-rater reliability was calculated to assess the consistency and effectiveness of the two raters in applying the annotation guidelines.

Inter-rater reliability was calculated using standard metrics such as Cohen's Kappa and Krippendorff's alpha (K-alpha). Across the initial corpus, the K-alpha coefficient was 0.40, indicating a *moderate* level of consensus between the two annotators and suggesting they had a solid grasp of the guidelines. However, the process revealed areas for improvement, which were incorporated into later phases of the study. Specifically, three annotation challenges were identified and addressed: (i) incomplete tags (e.g., an opening tag without a corresponding closing tag), (ii) nested or overlapping tags, which required more precise criteria, such as limiting the number of tags per feature to two; and (iii) tagging inconsistencies where a tag was correctly applied, but the suggested correction did not align with the feature's intent. The initial annotations were conducted manually, but more user-friendly platforms were later explored, leading to the adoption of LABELBOX[55]. This annotation tool was selected as the primary web-based annotation tool for subsequent studies. It is known to facilitate the creation of high-quality annotated datasets and to streamline the labeling process (Colucci Cante et al. (2024)).

The resulting annotated corpus from (Da Corte and Baptista (2024a)) is one of the few DevEd-specific writing corpora that integrates human annotation[56]. This resource represents a significant contribution to the field of Applied Linguistics, particularly in advancing resource-building efforts in corpus linguistics and Natural Language Processing (NLP) for educational purposes. The corpus was later expanded from 100 to 300 (Da Corte and Baptista (2025c)), just with texts deemed as DevEd by ACCUPLACER, and then (Da Corte and Baptista (2025b)), from 300 to 450, incorporating an additional 150 texts from students placed at the College Level. As mentioned before, college-level is the denomination used to represent the proficiency benchmark required to participate in an academic program. The progressive expansion of the corpus provided a balanced dataset for comparative analysis and classification, and is addressed throughout this conclusion.

As part of the annotation procedure developed for this study, the task extended beyond simply marking errors in student writing to include the identification of higher-order features,

[55] https://docs.labelbox.com/

[56] The full datasets accompanying this annotated corpus will be made publicly available after the thesis defense. Selected datasets were included either in key sections of this document or in the published papers presented.

most notably, multiword expressions (MWE), which often signal native language proficiency (Savary et al. (2019); Kochmar et al. (2020); Hinkel (2022)). Motivated by the need to offer more accurate linguistic insights about the writing of DevEd students, in (Da Corte and Baptista (2022)) a phraseological framework for MWE analysis was introduced and tested, examining its relationship (and value) to both writing proficiency and placement accuracy. This framework was applied to a corpus of 186 texts (86 additional texts beyond the 100 from the original corpus) using a focused MWE annotation protocol, and several Machine Learning (ML) models were trained and evaluated. Results demonstrated promising performance gains, particularly for the Neural Network (NN) algorithm, which showed an 8.1% increase in classification accuracy (from 74.2% to 82.3%) due to the inclusion of MWE features.

This improvement raised an important question: *Could the annotation framework be extended cross-linguistically*? Although this thesis focused on English (L1), TCC offers courses in multiple languages, with Spanish being one of the most commonly studied second languages. The English-based annotation framework in (Da Corte and Baptista (2022)) was therefore extended and applied to a new corpus assembled with texts from a *Spanish Grammar and Composition* course, produced within (and throughout the duration of) this course, which was the primary focus in (Da Corte and Baptista (2023)). This course serves as both a preparatory and a remedial course for native and non-native Spanish speakers pursuing a Spanish Translation certificate. The corpus consisted of 41 essays: 16 written by native Spanish speakers and 25 by non-native Spanish speakers participating in the course. Given the developmental nature of the course, adapting the framework to Spanish provided comparative linguistic insights and enabled further exploration of MWE as potential indicators of writing proficiency. This cross-linguistic application tested and confirmed, through ML techniques, that MWE, particularly *verbal idioms* and *compound nouns*, can serve as reliable linguistic markers of proficiency and contribute to a more accurate classification of skill level.

While both the Spanish and English corpora confirmed the value of MWE as linguistic markers of proficiency, the English (L1) corpus, annotated more extensively with 21 linguistically grounded features, allowed for a more fine-grained analysis of developmental writing. Beyond phraseological patterns, this annotation process shed some light on how writing evolves and which features distinguish one proficiency level from another. For instance, texts at DevEd Level 1 tended to be shorter (averaging 197 words), displayed more frequent foundational issues such as punctuation (PUNCT), orthographic errors (ORT), and often lacked examples or supporting statements involving reasoning (EXAMPLE; REASON). In contrast, DevEd Level 2 texts were typically longer (averaging 335 words), showed a marked reduction in lower-order errors (PUNCT; ORT), and included more higher-order features such as sentence variety, rhetorical clarity, and discourse-level cohesion. These texts also showed more consistent use of examples and reasoning[57], aligning with findings that as proficiency increases, students move

[57] **Appendix D** showcases two sample annotated texts, one from DevEd Level 1 and one from DevEd Level 2, for comparison.

from correcting basic errors to refining clarity and expression (Nygård and Hundal (2024)). This progression highlights the value of fine-grained linguistic features in capturing developmental trajectories, which, ultimately, can support more precise and targeted placement.

Although this thesis focused on a community college in Oklahoma, the framework and annotation protocols developed are both scalable and adaptable to other institutions, particularly those with similar placement systems and programs that support students in developing their writing skills. According to data from the American Association of Community Colleges (AACC)[58], there are 1,024 community colleges across the U.S., serving approximately 39% of the nation's undergraduate student population (American Association of Community Colleges (2025)). These statistics highlight the relevance and potential impact of the framework (and its corresponding methodology) presented in this study. More broadly, this thesis contributes to the field of Applied Linguistics by defining and systematizing a taxonomy of features relevant to early-stage academic writing in native English, grounded in authentic, manually annotated student texts.

Given the broader applicability of the proposed framework and its potential to inform more accurate assessment and placement practices, it is essential to revisit the second and third challenges outlined earlier, namely, (ii) the lack of transparency and linguistic grounding in ACCUPLACER's placement process, and (iii) the high and concerning misplacement rate associated with such system. These concerns motivated a series of studies (Da Corte and Baptista (2024b, 2025d)) that sought to model and examine ACCUPLACER's classification methodology, particularly because its placement algorithm, as well as the features factored into the assessment process, if any, are not publicly available. Since language assessment and placement were at the core of this thesis, an examination of this nature was necessary to understand how automated systems evaluate writing and to identify opportunities to improve both accuracy and transparency.

First, a methodology had to be devised to attempt to approximate ACCUPLACER's internal logic. Several experiments were designed to replicate its placement decisions, as described in (Da Corte and Baptista (2025d)). In preparation, each of the 100 essays from the original corpus, already annotated with DevEd-specific features, was also assessed by skill level: DevEd Level 1, DevEd Level 2, and College Level. These level assignments were based on slight adaptations of the official institutional DevEd course descriptions. Notably, the skill-level classifications relied not on the linguistic features developed in this thesis, but on the same high-order, broadly defined textual descriptors used by ACCUPLACER. Because these descriptors were accompanied by an 8-point Likert scale ACCUPLACER (The College Board, 2022, p. 26), which could be challenging to operationalize reliably, a simplified 4-point Likert scale was created using simpler, more intuitive language that aligned with the DevEd context. The 4-point labels (and their numerical values), *deficient* = 0, *below average* = 1, *above average* = 2, and *outstanding* = 3, referred to

[58] https://www.aacc.nche.edu/

the overall quality of the texts. While the descriptors themselves remained consistent with what was included in ACCUPLACER's manual, the modified scale was designed to help reduce potential misinterpretation and promote consistency across raters.

As expected, and despite the small size of the corpus, only about one-third of the essays received identical scores across all descriptors. The remaining two-thirds showed score discrepancies in one or more descriptors, confirming interpretive challenges with the definitions and a possible decline in scoring consistency as the number of descriptors increased. In response to these results, two descriptors, namely *mechanical conventions* and *organization & structure*, were enhanced by the expert linguists who developed the DevEd feature framework. These enhancements introduced more measurable elements, such as a numerical scale for counting typos under *mechanical conventions*, and clear, easily identifiable components of paragraph structure under *organization & structure*, including introduction, thesis, supporting statements, and conclusion. Inter-rater reliability scores were notably higher for these two descriptors, likely due to their clearer operationalization. By refining the scope of broad descriptors into specific, quantifiable features, these enhancements contributed to improved scale design and greater scoring consistency.

Conversely, the results for skill-level classifications showed the opposite pattern: about two-thirds of the essays received the same skill-level score, while one-third had differing ratings. To resolve these discrepancies in both descriptor and level-based assessments, a third rater was brought in to evaluate all texts following the same guidelines and protocols. Bringing a third rater to the assessment not only added methodological rigor to the process but also enabled the creation of a gold standard. This gold standard, made publicly available via (Da Corte and Baptista (2025d)), represents another significant contribution of this thesis and offers a validated reference point for assessing DevEd writing, a resource that, to the best of one's knowledge, has not previously existed.

In this gold standard, human raters classified 72 essays as DevEd Level 1 and 28 as DevEd Level 2, while ACCUPLACER classified an even 50 texts in each level. When these results were visualized in a confusion matrix, ACCUPLACER demonstrated a 22% misplacement rate relative to the human classifications and yielded a *weak* Pearson correlation (ρ = 0.22). In practical terms, this means that one in five students were placed into DevEd Level 2 by ACCUPLACER despite still needing support to develop their writing skills. This result compares to findings in the literature (Ganga and Mazzariello (2019)) regarding the misplacement rates often associated with automated placement systems. As argued throughout this thesis, the ultimate goal is not to discard automated systems, but to enhance them to support a more systematic, linguistically informed approach to student placement.

Leveraging this gold standard, a series of ML experiments were conducted in the same study (Da Corte and Baptista (2025d)) using ORANGE. ORANGE, as it has been addressed, is a free and user-friendly data-mining toolkit that provides a wide range of ML algorithms suitable for classification tasks such as the one at hand. The purpose of these experiments was twofold:

(i) to evaluate ACCUPLACER's performance in terms of accuracy relative to human classifications, without building an entirely new placement model, and (ii) to further investigate the limitations of current automated systems in supporting accurate student placement.

To carry out these experiments, several known ML models were tested, namely, Decision Tree (DT), Naïve Bayes (NB), Neural Network (NN), and Random Forest (RF). Given the small sample size, controlling for overfitting was essential. As a safeguard, the dataset was balanced by level and configured for 3-fold cross-validation, with two-thirds of the data reserved for training and one-third for testing. Due to this configuration, stratified cross-validation was not feasible. Among the tested models, RF and NN yielded the highest accuracy scores (75.7%). This high classification accuracy result offered a dual interpretation: (i) when relying on linguistic descriptors and human-annotated scores to replicate ACCUPLACER's classification task, approximately three out of ten students are still inaccurately placed in DevEd courses; and (ii) although the official descriptors were broad and abstract, when enhanced with human input, they contributed meaningfully to improved placement accuracy.

ML techniques also allowed the ranking of the six descriptors using Information Gain and Chi-square ranking methods. Because the Pearson correlation coefficient between the two ranking methods was considered *very strong*, ($\rho = 0.824$) the Chi-square method was adopted for subsequent experiments. This process helped identify which descriptors were most predictive of writing proficiency and placement. Notably, the two descriptors that were enhanced, *mechanical conventions* and *organization & structure*, were ranked among the top four. The other two were *development & support* and *sentence variety & style*. When additional classification experiments were conducted using only the top four descriptors, the NB model achieved an accuracy of 81.1%, followed by the NN, which maintained its previous score of 75.7%. NB and RF are models that have been recommended in the classic ML literature for similar tasks, not only for their predictive power and interpretability but also for their pedagogical utility (Huang (2023); Pajila et al. (2023)).

These results confirmed the value of human input in descriptor enhancement. They highlighted a promising direction for future work, not only to continue refining the remaining four descriptors but also to explore other linguistic features that may more accurately capture the skill level of students at the onset of their college careers. Beyond understanding the value of human input in descriptor refinement, leveraging ML techniques in this study also presented actionable opportunities for higher education institutions, and more importantly, for faculty and curriculum designers, to more effectively target instruction and interventions toward areas that directly impact placement outcomes.

Due to the need to explore additional linguistic features that may better capture the skill level of students as they start college, and, thereby, support a more accurate assessment of writing proficiency and college readiness, this next study (Da Corte and Baptista (2024c)) took a step further to expand on the identification of linguistic features critical for DevEd placement

decisions, and to see if these features could enhance students' English (L1) writing proficiency assessment within DevEd.

To achieve these goals, two NLP tools, Coh-Metrix and CTAP, were used to mechanically extract predefined (textual, syntactic, lexical, and discourse-related) linguistic features by parsing each of the 100 texts, again, from the original corpus. A total of 445 features were extracted (106 from Coh-Metrix and 339 from CTAP), each expressed by an attribute-value pair. Due to the extensive list of features, some general examples are provided here. The features extracted from Coh-Metrix were *mean number of words per sentence* (syntactic complexity); *measure of textual lexical diversity [MTLD]* (lexical diversity); *type-token ratio [TTR]* (vocabulary use); *causal and temporal connectives* (cohesion); *word frequency norms* (lexical sophistication); *narrativity and situation model indices* (discourse-level coherence). Regarding CTAP, features extracted included *Parts of Speech (POS) tag distributions* (e.g., noun/verb density); *syntactic parse depth and dependency lengths*; *number of subordinate clauses*; *prepositional phrases per sentence*; *sophistication indices using NGSL*[59]*-based metrics*.

These features are standardized across corpora, enabling consistent comparisons and supporting the identification of key differences in writing proficiency. As such, they could help distinguish developmental-level texts from more advanced writing, while also adding depth to the linguistic analysis through computational precision. These mechanically-derived features were paired with the 21 DevEd-specific (DES) features, which, by contrast, were devised based on human observation, anchored in the literature, and manually annotated. This pairing yielded a 457-feature set that combines the consistency of automated extraction with the pedagogical insight of human annotation, representing another key contribution of this thesis toward improving placement accuracy beyond what Accuplacer currently achieves.

While there may be partial *conceptual* overlap between features extracted by Coh-Metrix and CTAP (e.g., sentence length, word frequency), such overlap does not extend to the DES features. DES constructs are context-bound and, apparently, have not yet been considered within the automatically extracted feature sets of CTAP or Coh-Metrix. As such, DES features still require manual annotation. Future work entails operationalizing these features in ways that would enable their automatic detection and extraction. Achieving this goal will require extensive research and collaboration with NLP experts to develop strategies that can approximate each DES construct from a human-defined linguistic concept into a computational representation that NLP algorithms can reliably detect. These strategies will likely differ for each DES feature, as they are designed to measure a range of distinct linguistic aspects, some capturing deviations from proficiency, others reflecting academic writing standards. Additionally, as shown in **Appendix D**, DES features do not appear uniformly across texts; their presence and distribution vary both across individual texts and by proficiency levels (DevEd Levels 1 and 2, and College Level).

[59] As noted in the list of abbreviations, NGSL stands for New General Service List, a free online text analysis tool with wordlists designed for English as a Second or Foreign Language learners.

To evaluate the effectiveness and broader applicability of the comprehensive 457-feature set developed in this same study, four supervised ML experiments were conducted using ORANGE, across two classification scenarios: (a) full-length essays and (b) split samples (100 words). The inclusion of split essays provided an additional analytical layer to assess whether text length influenced classification performance or feature ranking, with implications for practical placement settings.

A total of ten ML algorithms were tested, including previously explored models such as DT, NB, NN, and RF, as part of the broader analysis in this study (Da Corte and Baptista (2024c)). The experiments pursued two key objectives: (i) identifying the most predictive linguistic features and (ii) determining the most effective algorithms for writing placement classification. Each experiment offered distinct contributions: (1) Experiment 1 used each feature set from COH-METRIX, CTAP, and DES to classify text samples and establish a baseline, providing a previously unavailable point of comparison and, thus, constituting a contribution in itself; (2) Experiment 2 evaluated performance using the top-ranked 11 features (from each source) based on Information Gain and Chi-square scoring; (3) Experiment 3 employed a leave-one-cluster-out strategy to assess the unique contribution of each feature cluster, grouping lexical, semantic, and discursive features by source (i.e., COH-METRIX, CTAP, or DES) without cross-source mixing; and (4) Experiment 4 combined features across COH-METRIX, CTAP, and DES, ranking them with Information Gain in iterative packs of ten until ML models' performance plateaued.

In summary, the baseline classification accuracy (CA) in Experiment 1 was 72.7%, achieved by the RF model using full-text samples. Notably, Experiment 2 demonstrated significant gains when models were trained using only the top 11 features identified through Information Gain. CA increased by nearly 14% (to 78.8%) for NB with top-ranked features from COH-METRIX. While gains using DES features were more modest (e.g., a 6.1% increase for Decision Tree (DT); CA < 60%), several DevEd-specific features, such as EXAMPLE, REASON, MWE, ORT, and PUNCT, consistently ranked among the top predictors. In Experiment 3, applying a leave-one-cluster-out strategy revealed that removing specific feature clusters significantly hindered performance, particularly for the k-Nearest Neighbor (kNN), LR, and CN2 Rule Induction (CN2) models, confirming the complementary value of diverse linguistic feature types.

Experiment 4 was the most comprehensive of all, incorporating ranked features from all three feature sources. This experiment focused exclusively on full-text scenarios, given their consistent superiority across prior experiments. Here, the NB model achieved the highest overall accuracy, with a 9.1% increase over the baseline (72.7%), reaching a CA score of 81.8%. Among the top 60 ranked features, most originated from CTAP, but novel DES features, notably VAGREE (7^{th}) and MWE (14^{th}), ranked within the top 15, confirming their value in capturing developmental writing. Beyond the statistical gains and their implications for CA, these results reinforce a central argument of this thesis: placement practices can and should integrate interpretable, human-derived linguistic features to support more transparent and pedagogically

sound decision-making, especially at the critical entry point of a student's higher education journey. These findings further emphasized the importance of lexical and syntactic complexity analysis in improving placement accuracy (Stavans and Zadunaisky-Ehrlich (2024)).

To further evaluate the effectiveness and potential to support fairer, more transparent placement practices, (Da Corte and Baptista (2025c)) extended the original experimental setup from (Da Corte and Baptista (2024c)) and applied it to an expanded corpus of 300 essays (previously, 100). The additional 200 essays were collected using the same stratified random sampling procedures as in previous experiments and maintained balance across DevEd levels. This larger corpus allowed for a more robust test of the generalizability of the 457 linguistic features identified in earlier experiments. The 200 essays were annotated by the same two human raters previously mentioned, who also assigned a skill level to each text. An agreement was reached between the raters on 260 texts, with 40 discrepancies resolved according to the established skill-level classification procedures. This step was essential for generating a reliable reference point for comparison with ACCUPLACER classifications, now tested against an expanded corpus.

Comparisons between human-assigned and ACCUPLACER-generated levels were conducted separately for the original 100 texts, the newly added 200, and the combined corpus of 300. Across all three comparisons, Pearson correlation coefficients remained *weak* (ρ = 0.304, on average), reaffirming ACCUPLACER's limitations in aligning with human judgments across proficiency levels. On the full corpus, ACCUPLACER correctly matched the human-assigned level in 236 out of 300 texts, resulting in an accuracy score of 79%. This score indicated that approximately one in five students is still placed at an incorrect level, emphasizing the pedagogical consequences these misclassifications have on learning outcomes and student success.

Importantly, in (Da Corte and Baptista (2025c)), the experimental setup from (Da Corte and Baptista (2024c)) was revisited, using this expanded corpus, to also assess whether the classification performance and feature rankings of ML algorithms remained consistent. The study continued to rely on classical supervised ML models. Using Information Gain as the ranking method, the original 457-feature set was refined, reducing the 21 DES features to 11 based on their predictive power for placement accuracy. The goal was to determine whether the previously top-ranked features retained their predictive value and whether the CA of the ML models improved with the updated feature set.

The ML experiments, using ORANGE, were set in two scenarios: (a) Scenario 1, using the full corpus with classification levels automatically assigned by ACCUPLACER; and (b) Scenario 2, using the same corpus with classification levels assigned by human annotators. In each scenario, experiments were first conducted with the refined feature set, now comprising 445 features, followed by additional experiments where features were added incrementally in bundles of ten until classification results reached asymptotic results.

When all features were combined in Scenario 1, seven out of the ten ML models tested surpassed the baseline CA of 72.7% established in (Da Corte and Baptista (2024c)), with results

ranging from 75% using RF to 80% using GB. These improvements confirmed the potential of the refined feature set to enhance placement accuracy. Notably, the highest-ranked features were drawn predominantly from CTAP (99%), followed by two Coh-Metrix features, namely, word count, number of words (35th) and sentence count, number of sentences (88th), and one DES feature, VAGREE (117th), which had previously ranked 7th in earlier experiments conducted on a smaller corpus of 100 texts.

After features were grouped into sets of ten and bundles were tested iteratively to identify optimal classification performance before reaching asymptotic results, the NB model showed strong and consistent results, maintaining CA scores of 77% or higher and plateauing after 30 features were included. At this 30-feature peak, the impact of DES features was further tested by supplementing the set of 30 with all 11 DES features, yielding a total of 41. This addition resulted in CA gains for both Support Vector Machine (SVM) (81%) and NN (80.5%), improving by 3% and 4.5%, respectively. These gains confirmed the value of DES features as meaningful indicators of students' writing ability.

In Scenario 2, which used the same corpus but relied on human-assigned classification levels, no models surpassed the 72.7% CA baseline. However, some shifts in feature rankings were observed, reflecting differences in how human raters and automated systems interpret linguistic patterns. The majority of top-ranked features still came from CTAP (67.5%), followed by Coh-Metrix (11%), with two DES features, EXAMPLE and ORT, ranking 9th and 31st, respectively. These findings reaffirmed that DES features continue to offer valuable insights into writing performance across evaluation scenarios.

When features were tested in bundles of ten again, the LR model peaked at a CA of 78.5% (slightly surpassing the 72.7% baseline) with 40 features, after which its performance declined. Following this step, a combination of these top 40 features and the remaining nine DES features (two of which were already included in the top 40) was tested. The NB model matched LR's peak CA score of 78.5% in this configuration. This result further demonstrated (i) the potential of DES features to yield valuable insights into writing performance, and (ii) NB's effectiveness in pattern detection and feature selection, as supported in current literature (Hirokawa (2018)).

In addition to the insights gained from the linguistic features explored throughout this thesis, a sociodemographic study was designed, conducted, and published to ensure a well-rounded analysis, linking it to placement outcomes in the community college setting. Findings from (Poe et al. (2019); Jensen and Giordano (2022)) further informed the need for this additional study, noting that automatic placement outcomes often reflect a mismatch between individual student performance data and broader institutional demographics (Castillo and Ávila Reyes (2025)). This concern was amplified and explained by (Arnold et al. (2024); Nastal and Messer (2025)), who warned that certain demographic groups may be disproportionately affected by such outcomes. As one of the two final contributions from this thesis, (Da Corte and Baptista (2025a)) examined this gap by investigating the relationship between race and gender and the

linguistic features present in student writing, as well as how these sociodemographic variables may influence placement outcomes[60].

For the analysis proper, a subset of the 300-text corpus was used. The final sample, constituting 210 texts, consisted of those for which participants self-identified as native English (L1) speakers and provided complete demographic information on race and gender. Even with this smaller subsample, and as outlined in the introduction, the 210 participants who produced the writing samples reflected both the overall institutional demographics and those of the broader ACCUPLACER test-taking population. Drawing from the findings in (Da Corte and Baptista (2025c)), a combination of the top-ranked linguistic features selected from COH-METRIX, CTAP, and all 11 DES features was used, totaling 40 features. These were employed to investigate potential disparities across gender and race using three statistical tests: (i) one-way ANOVA; (ii) Tukey's HSD; and (iii) Chi-square.

Results from the ANOVA revealed that not all linguistic features are equally sensitive to demographic variation. Gender did not show a *statistically significant* influence over any of the features tested, suggesting that variation in writing performance may not be driven by gender. With race, a total of nine features, especially those tied to syntactic complexity, discourse markers, and lexical precision, revealed *significant* variation. Most of the features were found within the DES feature set: PUNCT, WORDOMIT, PRECISION, ORT, and REASON, suggesting that race-based differences may be more evident when using targeted, human-devised descriptors.

Tukey's HSD confirmed that, while ANOVA indicated some variation, none of the gender-based pairwise comparisons reached significance. Regarding race, however, significant differences were noted after running Tukey's HSD test. Black or African American students demonstrated higher rates of word omission and lower use of proper punctuation and lexical precision compared to other racial groups considered (White and American Indian). On the contrary, they exhibited a higher incidence of positive connectives despite having lower mechanical accuracy. The comparison between automated and human-assigned placement classifications, using the Chi-square test, showed *no significant* associations with gender. However, a *borderline significant* association between race and human-assigned placement levels was noted. This association could have been unintentionally influenced, in part, by the nature and definitions of certain DES features, raising questions about how rater judgments may interact with linguistic patterns that intersect with race-based dialect differences.

These findings prompted three key considerations for discussion, each addressing a different but complementary dimension of placement practices: (i) human evaluation, anchored in a sound linguistic framework and appropriate training, is needed as it offers insights into how students write, something that automated placement systems cannot replicate; (ii) DES features

[60]Preliminary results from this study were presented at the 1st International Conference on Global Perspectives on Technology-Enhanced Language Learning and Translation (GLOTECH), held at the University of Alicante, 12–13 December 2024. The abstract is available in the conference's book of abstracts.

were carefully designed to capture the linguistic patterns of developing writers and remain a valuable resource for targeted feedback and pedagogically relevant analysis, particularly within the DevEd context; and (iii) equitable placement must be paired with instruction that is both culturally responsive and linguistically informed (Suárez-Álvarez et al. (2023); Lyons et al. (2025); Matta and Hamsho (2025)). Building on these considerations, future investigations will focus on (i) replicating the statistical analysis with a larger dataset will be important to determine whether classification accuracy can be preserved or improved when potentially bias-sensitive features are excluded or replaced; (ii) using other DES features, such as EXAMPLE, MWE (multiword expressions), and VAGREE (verbal agreement), not associated initially with race-based differences, could still be leveraged to potentially uncover valuable insights into student writing without reproducing existing disparities; and (iii) enhancing rater training to deepen understanding of how linguistic patterns intersect with race-based dialect differences, ensuring placement decisions remain fair and contextually informed.

While this thesis has offered methodologically sound and linguistically grounded evidence to refine placement practices through corpus development and feature-based analysis, it also opened opportunities to explore complementary approaches that could extend these insights into emerging technological frameworks, especially LLMs.

This investigation was conducted during a period of rapid scientific and technological development, not only in the U.S. but worldwide, marked by the breakthrough of newer LLMs, such as OpenAI ChatGPT, as noted earlier. Given this context, it felt like a natural progression to connect the assessment of students' writing skills with LLMs, exploring whether these models could enhance current placement processes and help reimagine placement practices in higher education based on their potential benefits (Crompton and Burke (2023); U.S. Department of Education, Office of Educational Technology (2023)).

In this light, the last study in this thesis (Da Corte and Baptista (2025b)) examined the potential of LLMs as more transparent, prompt-customizable, and complementary to existing classification systems, such as ACCUPLACER. Four well-known models, namely, Claude, Gemini, ChatGPT-3.5-turbo, and ChatGPT-4o, were evaluated for their ability to classify texts into DevEd Level 1, DevEd Level 2, and College level, with performance compared against both a human-rated gold standard and ACCUPLACER.

To support a more transparent and linguistically grounded placement process, two classification experiments were devised: (1) a one-step classification, in which each LLM classified texts as College or non-College, in one run; a (2) a two-step classification, in which, the same four LLMs reclassified the texts labeled as non-College (in the one-step classification) into the two known DevEd Levels: DevEd Level 1 or DevEd Level 2. Each experiment applied a four-prompting condition framework: (a) Zero-shot, (b) Few-shot, (c) Enhanced, and (d) Enhanced+, enabling a comprehensive analysis of whether classification precision improved when prompts were enriched with linguistic details and writing samples. The prompts were informed

by current literature and designed to reflect linguistic markers of developing writers in DevEd contexts. They represent an additional contribution of this thesis and can be refined or adapted for similar classification tasks.

In preparation for the classification tasks, additional text samples were incorporated into the existing 300-text corpus assembled in (Da Corte and Baptista (2025c)), following the same protocols used in earlier experiments. Because this task required college-level texts, 150 samples automatically classified as such by ACCUPLACER were randomly selected from the same exam database. The final corpus now consisted of 450 texts, evenly distributed across three levels: 150 each for DevEd Level 1, DevEd Level 2, and College level. In addition to ACCUPLACER's assessment, all essays were also independently evaluated by the two trained raters mentioned throughout this investigation.

Overall, all LLMs showed potential across classification scenarios, with Claude and Gemini consistently achieving higher precision as prompts became more specific (Claude: 62.22% in the Enhanced+ strategy, 1-step classification; Gemini: 69.33% in the Enhanced+ strategy, 2-step classification). Precision was selected as the primary evaluation metric because the goal was to ensure that the models' positive predictions were reliable, which is an essential requirement for DevEd/College-level placement. Future work will include the exploration of secondary metrics, such as Recall and F1 Score, to better differentiate between models that exhibit similar precision across different prompting strategies.

Having demonstrated the LLMs' potential in terms of precision, the analysis next examined how closely each LLM's classifications aligned with those produced by humans. Claude consistently achieved the highest Pearson correlation with the human-rated gold standard (ρ = 0.750; *substantial* agreement). This statistical measure was used to assess the alignment of LLM-generated classifications with human judgments, and Claude's performance exceeded that of ACCUPLACER (ρ = 0.670) in this regard. These findings also correspond with recent studies advocating for the use of LLMs in writing assessment and placement decisions, where they can contribute to increased fairness and transparency in high-stakes educational settings (Seßler et al. (2025); Wang et al. (2024b)). Additionally, early results from (Da Corte and Baptista (2025b)) suggest that certain LLMs may offer linguistically more robust feedback[61] than ACCUPLACER, providing preliminary insights into students' writing strengths and areas for improvement.

Further research is needed to: (i) evaluate the accuracy of the feedback these models generate; (ii) how the feedback is influenced by prompt design; and (iii) how effectively the feedback captures the key linguistic features that distinguish DevEd from college-level writing. Investigating these questions, while accounting for the evolving capabilities of LLMs, the non-deterministic nature of their outputs, and the critical role of prompt engineering and bias auditing, represents a clear and necessary direction for future work that will take place beyond the scope of this thesis.

[61] For illustration purposes, samples of the feedback generated from both Gemini and Claude are detailed in **Appendix I**.

What began as a linguistic and educational question of how to more accurately and equitably assess the writing of community college students was addressed through a series of studies that contributed to raising awareness towards DevEd students' writing and their representation (or lack thereof) in established corpora. While the design of this thesis did not aim to compete with existing automated classification tools, more precisely ACCUPLACER, it exposed important issues of transparency and limited linguistic grounding in their placement processes. Over the past three years, a series of experiments has confirmed the high and concerning rate of misplacement of the classification system analyzed in this thesis, raising important questions about how writing proficiency is currently assessed.

The findings from this thesis, from its early stages, have been shared with the scientific community, faculty, and key stakeholders at TCC, who are responsible for assessment and placement policies. The goal has not been to prescribe or dictate specific policy changes but to raise awareness of the limitations of current practices that rely primarily on automated systems, and to encourage informed conversations around placement accuracy and adequate student support. Ultimately, the results of this thesis aim to create a space for critical reflection and discussion among faculty and administrators on how placement decisions are made and what they should entail. While automatic placement misclassifications cannot be entirely avoided, as they result from algorithmic processes that operate independently of human judgment, faculty and administrators can play an active role in developing strategies to support students' linguistic development, ensuring they are well-prepared to succeed in their academic programs and beyond.

References

Abba, K., Zhang, S., and Joshi, R. (2018). Community college writers' metaknowledge of effective writing. *Journal of Writing Research*, 10(1):85–105. Available at: https://doi.org/10.17239/jowr-2018.10.01.04.

Abdi, H. and Williams, L. J. (2010). Tukey's honestly significant difference (hsd) test. In Salkind, N. J., editor, *Encyclopedia of Research Design*, volume 3, pages 1–5. SAGE Publications. Available at: https://personal.utdallas.edu/~Herve/abdi-HSD2010-pretty.pdf.

Adiguzel, T., Kaya, M. H., and Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, 15(3):1–13; ep429. Available at: https://doi.org/10.30935/cedtech/13152.

Akef, S., Mendes, A., Meurers, D., and Rebuschat, P. (2024). Investigating the generalizability of Portuguese readability assessment models trained using linguistic complexity features. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 332–341, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics. Available at https://aclanthology.org/2024.propor-1.34/.

American Association of Community Colleges (2025). Fast facts about community colleges in the United States 2025. https://www.aacc.nche.edu/research-trends/fast-facts/. American Association of Community Colleges, Washington, D.C.

Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic. Model card, available at https://www.anthropic.com/claude-3-model-card.

Arnold, L. R., Jiang, L., and Hassel, H. (2024). After implementation: Assessing student self-placement in college writing programs. *Journal of Writing Assessment*, 17(1):1–26. Available at: https://doi.org/10.5070/W4jwa.1571.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using Random Forests. *Research in Higher Education*, 60:1048–1064. Available at: https://doi.org/10.1007/s11162-019-09546-y.

Bickerstaff, S., Beal, K., Raufman, J., Lewy, E. B., and Slaughter, A. (2022). Five principles for reforming Developmental Education: A review of the evidence. Technical report, Center for the Analysis of Postsecondary Readiness. Available at: https://files.eric.ed.gov/fulltext/ED624623.pdf.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., and Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using Machine Learning. *Institute of Electrical and Electronics Engineers (IEEE) Access*, 9:95608–95621. Available at: https://doi.org/10.1109/ACCESS.2021.3093563.

Camardelle, A., Kennedy, B., and Nalley, J. (2022). The state of Black students at community colleges. Research brief, Joint Center for Political and Economic Studies, Washington, D.C. Available at: https://jointcenter.org/wp-content/uploads/2022/09/The-State-of-Black-Students-at-Community-Colleges.pdf.

Castillo, C. and Ávila Reyes, N. (2025). Students' sociodemographic characteristics and writing performance: A systematic literature review. *Reading and Writing*, pages 1–37. Available at: https://doi.org/10.1007/s11145-025-10658-4.

Chen, L., Chen, P., and Lin, Z. (2020). Artificial Intelligence in Education: A review. *Institute of Electrical and Electronics Engineers (IEEE) Access*, 8:75264–75278. Available at: https://doi.org/10.1109/ACCESS.2020.2988510.

Chen, X. and Meurers, D. (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. Available at: https://aclanthology.org/W16-4113.

Colucci Cante, L., D'Angelo, S., Di Martino, B., and Graziano, M. (2024). Text annotation tools: A comprehensive review and comparative analysis. In *International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 353–362. Springer. Available at: https://dblp.org/rec/conf/cisis/CanteMG24.

Cormier, M. and Bickerstaff, S. (2019). Research on Developmental Education Instruction for Adult Literacy Learners. In Perin, D., editor, *The Wiley Handbook of Adult Literacy*, pages 541–561. John Wiley & Sons, Inc., Hoboken, NJ. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119261407.ch25.

Crompton, H. and Burke, D. (2023). Artificial Intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1):22. Available at: https://doi.org/10.1186/s41239-023-00392-8.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443. Available at: https://doi.org/10.17239/jowr-2020.11.03.01.

Crossley, S. A., Russell, D. R., Kyle, K., and Römer, U. (2017). Applying Natural Language Processing tools to a student academic writing corpus: How large are disciplinary differences across Science and Engineering fields? *Journal of Writing Analytics*, pages 48–81. Available at: https://wac.colostate.edu/docs/jwa/vol1/crossley.pdf.

Da Corte, M. and Baptista, J. (2022). A Phraseology Approach in Developmental Education Placement. In *Proceedings of Computational and Corpus-based Phraseology, EUROPHRAS 2022, Malaga, Spain*, pages 79–86. Available at: https://europhras.com/2022/wp-content/uploads/2022/09/draft-proceedings-europhras.pdf.

Da Corte, M. and Baptista, J. (2023). Multiword Expression Tagging of Spanish Native and Non-Native Speakers' Written Essays in a Grammar and Composition Developmental Course. *Romanica Olomucensia*, 35(1):23–40. Available at: https://romanica.upol.cz/artkey/rom-202301-0003.php.

Da Corte, M. and Baptista, J. (2024a). Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3046–3056, Torino, Italia. ELRA and ICCL. Available at: https://aclanthology.org/2024.lrec-main.272/.

Da Corte, M. and Baptista, J. (2024b). Enhancing writing proficiency classification in Developmental Education: The quest for accuracy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6134–6143, Torino, Italia. ELRA and ICCL. Available at: https://aclanthology.org/2024.lrec-main.542/.

Da Corte, M. and Baptista, J. (2024c). Leveraging NLP and Machine Learning for English (L1) writing assessment in Developmental Education. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2-4 May, 2024, Angers, France*, volume 2, pages 128–140. Available at: https://www.scitepress.org/Papers/2024/127405/127405.pdf.

Da Corte, M. and Baptista, J. (2025a). Beyond the score: Exploring the intersection between sociodemographics and linguistic features in English (L1) writing placement. In Baptista, J. and Barateiro, J., editors, *14th Symposium on Languages, Applications and Technologies (SLATE 2025)*, volume 135 of *Open Access Series in Informatics (OASIcs)*, pages 6:1–6:18, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Available at: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2025.6.

Da Corte, M. and Baptista, J. (2025b). From prediction to precision: Leveraging LLMs for equitable and data-driven writing placement in Developmental Education. In Baptista, J. and Barateiro, J., editors, *14th Symposium on Languages, Applications and Technologies (SLATE 2025)*, volume 135 of *Open Access Series in Informatics (OASIcs)*, pages 1:1–1:18, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Available at: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2025.1.

Da Corte, M. and Baptista, J. (2025c). Refining English writing proficiency assessment and placement in Developmental Education using NLP tools and Machine Learning. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 288–303. INSTICC, SciTePress. Available at: https://www.scitepress.org/PublishedPapers/2025/133515/.

Da Corte, M. and Baptista, J. (2025d). Toward consistency in writing proficiency assessment: Mitigating classification variability in Developmental Education. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU*, pages 139–150. INSTICC, SciTePress. Available at: https://www.scitepress.org/PublishedPapers/2025/133539/.

Dahunsi, T. N. and Ewata, T. O. (2025). An exploration of the structural and colligational characteristics of lexical bundles in L1–L2 corpora for English language teaching. *Language Teaching Research*, 29(2):472–488. Available at: https://doi.org/10.1177/13621688211066572.

Darkenwald-DeCola, J. A. (2021). *'In College, I'm the One People Go To': Lessons from Successful Developmental Literacy Students About the Transition to College-Level Courses Across Disciplines*. PhD thesis, Rutgers The State University of New Jersey, School of Graduate Studies. Available at ProQuest: https://search.proquest.com/openview/17b77ce3a0b4e321220d2d1e7d7ed3a4/1.pdf?cbl=18750&diss=y&pq-origsite=gscholar.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/. Year coverage: 2008 to present.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353. Available at: http://jmlr.org/papers/v14/demsar13a.html.

Denison-Furness, J., Donohue, S. L., Hamlin, A., and Russell, T. (2022). Welcome/not welcome: From discouragement to empowerment in the writing placement process at Central Oregon Community College. In Nastal, J., Poe, M., and Toth, C., editors, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education*, pages 107–127. The WAC Clearinghouse/University Press of Colorado. Available at: https://wac.colostate.edu/books/practice/two-year/.

Dowell, N. and Kovanović, V. (2022). Modeling Educational Discourse with Natural Language Processing. In Lang, C., Wise, A. F., Merceron, A., Gašević, D., and Siemens, G., editors, *The Handbook of Learning Analytics*, chapter 11, pages 105–119. Society for Learning Analytics Research (SOLAR), Vancouver, Canada, 2nd edition. Available at: https://www.solaresearch.org/publications/hla-22/.

Durrant, P. (2022). *Corpus Linguistics for Writing Development: A Guide for Research*. Routledge, 1st edition. Available at https://doi.org/10.4324/9781003152682. ISBN: 9780367715786.

East, M. and Slomp, D. (2024). The ethical turn in writing assessment: How far have we come, and where do we still need to go? *Language Teaching*, 57(2):262–273. Available at: https://doi.org/10.1017/S0261444823000034.

Edgecombe, N. and Weiss, M. (2024). Promoting equity in Developmental Education reform: A conversation with Nikki Edgecombe and Michael Weiss. Podcast Episode in *Evidence First*. Available at: https://postsecondaryreadiness.org/promoting-equity-developmental-education-reform/.

Engsig, T. T., Johnstone, C. J., and Schuelka, M. J. (2025). Creating inclusive educational spaces: assessing assessment in a post-pandemic world. *International Journal of Inclusive Education*, 29(8):1371–1388. Available at: https://doi.org/10.1080/13603116.2023.2274107.

Esfandiari, R. and Ahmadi, M. (2023). Phraseological complexity and academic writing proficiency in abstracts authored by student and expert writers. *English Teaching & Learning*, 47(4):429–448. Available at: https://doi.org/10.1007/s42321-022-00118-5.

Fields, J., Chovanec, K., and Madiraju, P. (2024). A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *Institute of Electrical and Electronics Engineers (IEEE) Access*, 12:6518–6531. Available at: https://doi.org/10.1109/ACCESS.2024.3349952.

Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic text difficulty estimation using embeddings and Neural Networks. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 335–348. Springer. Available at: https://doi.org/10.1007/978-3-030-29736-7_25.

Freelon, D. (2013). Recal OIR: ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, 8(1):10–16. Available at: https://dfreelon.org/utils/recalfront/recal-oir/.

Ganga, E. and Mazzariello, A. (2019). Modernizing college course placement by using multiple measures. Policy brief, Education Commission of the States and the Center for the Analysis of Postsecondary Readiness (CAPR). Available at: https://postsecondaryreadiness.org/wp-content/uploads/2019/03/Modernizing_College_Course_Placement_by_Using_Multiple_Measures_Final.pdf.

Gere, A. R. (2019). *Developing Writers in Higher Education: A Longitudinal Study*. University of Michigan Press. Available at https://doi.org/10.3998/mpub.10079890; ISBN: 978-0-472-13124-2.

Gilman, H., Giordano, J. B., Hancock, N., Hassel, H., Henson, L., Hern, K., Nastal, J., and Toth, C. (2019). Two-year college writing placement as fairness. *Journal of Writing Assessment*, 12(1). Available at: https://escholarship.org/uc/item/4zv0r9b2.

Gilquin, G. and Granger, S. (2022). Using data-driven learning in language teaching. In O'Keeffe, A. and McCarthy, M. J., editors, *The Routledge Handbook of Corpus Linguistics*, pages 430–442. Routledge, 2 edition. Available at: https://doi.org/10.4324/9780367076399-30.

Google for Developers (2025). Classification: ROC and AUC. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Götz, S. and Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1):1–38. Available at: https://doi.org/10.1075/ijlcr.00039.got.

Hanks, P. (2008). Lexical patterns: From Hornby to Hunston and beyond. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII EURALEX International Congress*, volume 1, pages 89–129, Barcelona, Spain. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Available at: https://uwe-repository.worktribe.com/output/1018859/lexical-patterns-from-hornby-to-hunston-and-beyond-the-hornby-lecture.

Hilte, L., Daelemans, W., and Vandekerckhove, R. (2020). Lexical patterns in adolescents' online writing: the impact of age, gender, and education. *Written Communication*, 37(3):365–400. Available at: https://doi.org/10.1177/0741088320917921.

Hinkel, E. (2022). Teaching and learning multiword expressions. In Hinkel, E., editor, *Handbook of Practical Second Language Teaching and Learning*, pages 435–448. Routledge, 1 edition. Available at: https://www.elihinkel.org/downloads/ch31_MultiwordExpressions.pdf.

Hirokawa, S. (2018). Key attribute for predicting student academic performance. In *Proceedings of the 10th International Conference on Education Technology and Computers*, ICETC '18, page 308–313, New York, NY, USA. Association for Computing Machinery. Available at: https://dl.acm.org/doi/10.1145/3290511.3290576.

Hirsch, S., Smith, K., and Sorapure, M. (2024). Collaborative writing placement: Partnering with students in the placement process. *Journal of Writing Assessment*, 17(2). Available at: https://escholarship.org/uc/item/4b85378n.

Holschuh, J. P. (2019). College reading and studying: The complexity of academic literacy task demands. *Journal of Adolescent & Adult Literacy*, 62(6):599–604. Available at: https://ila.onlinelibrary.wiley.com/doi/10.1002/jaal.876.

Hornby, A. S. (1954). *A Guide to Patterns and Usage in English*. Oxford University Press. Accessed via Tulsa Community College Library. It is also available at: https://openlibrary.org/books/OL5245578M/Guide_to_patterns_and_usage_in_English; ISBN: 978-0194313162.

Huang, Z. (2023). An intelligent scoring system for English writing based on Artificial Intelligence and Machine Learning. *International Journal of System Assurance Engineering and Management*, 14:1–8. Available at https://doi.org/10.1007/s13198-023-02105-w.

Hughes, S. and Li, R. (2019). Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, 41:72–75. Available at: https://www.sciencedirect.com/science/article/pii/S1075293519300613.

Hunston, S. and Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing. Available at: https://benjamins.com/catalog/scl.4. ISBN 9781556193989.

Jensen, D. L. and Giordano, J. B. (2022). Afterword. Placement, equity, and the promise of democratic open-access education. In Toth, C. and Sullivan, P., editors, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education*, pages 279–286. WAC Clearinghouse / University Press of Colorado. Available at: https://wac.colostate.edu/docs/books/two-year/afterword.pdf.

Johnson, M. S., Liu, X., and McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3):338–361. Available at: https://doi.org/10.1111/jedm.12335.

Khan, M. A. (2019). New Ways of Using Corpora for Teaching Vocabulary and Writing in the ESL Classroom. *ORTESOL Journal*, 36:17–24. Available at: https://files.eric.ed.gov/fulltext/EJ1219789.pdf.

Kim, Y.-S. G., Harris, K. R., Goldstone, R., Camping, A., and Graham, S. (2025). The science of teaching reading is incomplete without the science of writing: A randomized control trial of integrated teaching of reading and writing. *Scientific Studies of Reading*, 29(1):32–54. Available at: https://doi.org/10.1080/10888438.2024.2380272.

Kim, Y.-S. G., Wolters, A., and won Lee, J. (2024). Reading and writing relations are not uniform: They differ by the linguistic grain size, developmental phase, and measurement. *Review of Educational Research*, 94(3):311–342. Available at: https://doi.org/10.3102/00346543231178830.

Kochmar, E., Gooding, S., and Shardlow, M. (2020). Detecting multiword expression type helps lexical complexity assessment. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno,

A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association. Available at: https://aclanthology.org/2020.lrec-1.545/.

Koo, T. and Li, M. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163. Available at: https://doi.org/10.1016/j.jcm.2016.02.012.

Kopko, E., Brathwaite, J., and Raufman, J. (2022). The next phase of placement reform: Moving toward equity-centered practice. Research brief, Center for the Analysis of Postsecondary Readiness. Available at: https://files.eric.ed.gov/fulltext/ED626145.pdf.

Kosiewicz, H., Morales, C., and Cortes, K. E. (2024). The "missing English learner" in higher education: How identification, assessment, and placement shape the educational outcomes of English learners in community colleges. In Perna, L. W., editor, *Higher Education: Handbook of Theory and Research: Volume 39*, pages 325–379. Springer Nature Switzerland, Cham. Available at: https://doi.org/10.1007/978-3-031-38077-8_7.

Kubiak, J. M. (2022). Putting Accuplacer in its place: Expanding evidence in placement reform at Jamestown Community College. In Giordano, J. B., Jensen, D. L., and Hassel, H., editors, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education*, pages 151–172. The WAC Clearinghouse and University Press of Colorado, Fort Collins, CO. Available at: https://wac.colostate.edu/docs/books/two-year/chapter6.pdf.

Laporte, E. (2018). Choosing features for classifying multiword expressions. In Sailer, M. and Markantonatou, S., editors, *Multiword expressions: In-sights from a multi-lingual perspective*, pages 143–186. Language Science Press, Berlin. Available at: https://zenodo.org/records/1182597.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resource and Evaluation*, 58(1):73–110. Available at: https://doi.org/10.1007/s10579-023-09693-w.

Link, S. and Koltovskaia, S. (2023). Automated scoring of writing. In Kruse, O., Rapp, C., Anson, C. M., Benetos, K., Cotos, E., Devitt, A., and Shibani, A., editors, *Digital Writing Technologies in Higher Education: Theory, Research, and Practice*, pages 333–345. Springer International Publishing, Cham. Available at: https://doi.org/10.1007/978-3-031-36033-6_21.

Lukwaro, E. A. E., Kalegele, K., and Nyambo, D. G. (2024). A review on NLP techniques and associated challenges in extracting features from education data. *International Journal of Computing and Digital Systems*, 16(1):961–979. Available at: https://doi.org/10.12785/ijcds/160170.

Lyons, S., Oliveri, M. E., and Poe, M. (2025). A framework for enacting equity aims in assessment use: A justice-oriented approach. In *Culturally Responsive Assessment in Classrooms and Large-Scale Contexts*, pages 88–105. Routledge, New York, NY. Available at: https://doi.org/10.4324/9781003392217-7.

Ma, H., Wang, J., and He, L. (2024). Linguistic features distinguishing students' writing ability aligned with CEFR levels. *Applied Linguistics*, 45(4):637–657. Available at: https://doi.org/10.1093/applin/amad054.

MacArthur, C. A. (2023). Postsecondary Developmental Education in writing: Issues and research. In Liu, X., Hebert, M., and Alves, R. A., editors, *The Hitchhiker's Guide to Writing Research: A Festschrift for Steve Graham*, pages 335–355. Springer International Publishing, Cham. Available at: https://doi.org/10.1007/978-3-031-36472-3_18.

Matta, M. and Hamsho, N. (2025). Consequences of response formats on racial and ethnic bias and fairness in writing assessments. *School Psychology Review*, 0(0):1–14. Available at: https://doi.org/10.1080/2372966X.2025.2462984.

Mazzariello, A., Ganga, E., and Edgecombe, N. (2018). Developmental Education: An introduction for policymakers. Policy brief, Education Commission of the States; Center for the Analysis of Postsecondary Readiness (CAPR). Available at: https://postsecondaryreadiness.org/wp-content/uploads/2018/02/developmental-education-introduction-policymakers.pdf.

McDonald, N., Johri, A., Ali, A., and Collier, A. H. (2025). Generative Artificial Intelligence in Higher Education: Evidence from an Analysis of Institutional Policies and Guidelines. *Computers in Human Behavior: Artificial Humans*, 3:100121. Available at: https://doi.org/10.1016/j.chbah.2025.100121.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge Applied Linguistics. Routledge, London and New York. ISBN: 9780415286237.

McNamara, D. and Graesser, A. C. (2011). Coh-Metrix: An automated tool for theoretical and applied Natural Language Processing. In *Applied Natural Language Processing*, pages 188–205. IGI Global. Available at: https://www.igi-global.com/gateway/chapter/61049.

Miaschi, A., Dell'Orletta, F., and Venturi, G. (2024). Evaluating Large Language Models via linguistic profiling. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2835–2848, Miami, Florida, USA. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.emnlp-main.166/.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A survey. *ACM Computing Surveys*, 56(2):1–40. Available at: https://doi.org/10.1145/3605943.

Morris-Compton, D. (2021). Intent to persist among African American community college students in remedial education. *Journal of Negro Education*, 90(4):496–507. Available at: https://www.jstor.org/stable/10.2307/48679125.

Murray, N. and Sharpling, G. (2019). What traits do academics value in student writing? Insights from a psychometric approach. *Assessment & Evaluation in Higher Education*, 44(3):489–500. Available at: https://doi.org/10.1080/02602938.2018.1521372.

Nam, D. and Park, K. (2020). *I will write about*: Investigating multiword expressions in prospective students' argumentative writing. *Plos one*, 15(12):e0242843. Available at: https://doi.org/10.1371/journal.pone.0242843.

Nastal, J. (2019). Beyond tradition: Writing placement, fairness, and success at a two-year college. *Journal of Writing Assessment*, 12(1). Available at: https://escholarship.org/uc/item/4wg8w0ng.

Nastal, J. and Messer, K. (2025). Afterword: Finding the right note in writing placement. *Journal of Writing Assessment*, 18(1). Available at: https://doi.org/10.5070/W4.jwa.47034.

Nieminen, J. H. (2025). The paradox of inclusive assessment. *Assessment & Evaluation in Higher Education*, 50(4):564–576. Available at: https://doi.org/10.1080/02602938.2024.2419604.

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., Lyden, S., Hassan, G. M., and Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 29(2):126–153. Available at: https://doi.org/10.1080/22054952.2024.2372154.

Nygård, M. and Hundal, A. K. (2024). Features of grammatical writing competence among early writers in a Norwegian school context. *Languages*, 9(1):29. Available at: https://doi.org/10.3390/languages9010029.

OECD (2023). Survey of adult skills (country note: United states). Technical report, Organisation for Economic Co-operation and Development (OECD), Paris. Available at: https://www.oecd.org/skills/piaac/United-States-country-note.pdf.

Okinina, N., Frey, J.-C., and Weiss, Z. (2020). CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7123–7131, Marseille, France. European Language Resources Association. Available at: https://aclanthology.org/2020.lrec-1.880/.

Pajila, P. B., Sheena, B. G., Gayathri, A., Aswini, J., Nalini, M., et al. (2023). A comprehensive survey on Naive Bayes algorithm: Advantages, limitations and applications. In *4th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1228–1234. Institute of Electrical and Electronics Engineers (IEEE). Available at: https://ieeexplore.ieee.org/document/10276274.

Pal, A. K. and Pal, S. (2013). Classification model of prediction for placement of students. *International Journal of Modern Education and Computer Science*, 5(11):49. Available at: https://doi.org/10.5815/ijmecs.2013.11.07.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics. Available at: https://aclanthology.org/2020.coling-main.296/.

Pastor, G. C. (2017). Collocational constructions in translated Spanish: What corpora reveal. In Mitkov, R., editor, *Computational and Corpus-Based Phraseology - Second International Conference, Europhras 2017, London, UK, November 13-14, 2017, Proceedings*, volume 10596 of *Lecture Notes in Computer Science*, pages 29–40. Springer. Available at: https://doi.org/10.1007/978-3-319-69805-2_3.

Phelps, D., Pickard, T., Mi, M., Gow-Smith, E., and Villavicencio, A. (2024). Sign of the times: Evaluating the use of Large Language Models for idiomaticity detection. In Bhatia, A., Bouma, G., Doğruöz, A. S., Evang, K., Garcia, M., Giouli, V., Han, L., Nivre, J., and Rademaker, A., editors, *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL. Available at: https://aclanthology.org/2024.mwe-1.22/.

Pichai, S. and Hassabis, D. (2023). Introducing Gemini: our largest and most capable AI model. Available at: https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini.

Pittman, R. T., O'Neal, L., Wright, K., and White, B. R. (2024). Elevating students' oral and written language: Empowering African American students through language. *Education Sciences*, 14(11):1191. Available at: https://www.mdpi.com/2227-7102/14/11/1191.

Poe, M., Nastal, J., and Elliot, N. (2019). Reflection. An admitted student is a qualified student: A roadmap for writing placement in the two-year college. *Journal of Writing Assessment*, 12(1). Available at: https://escholarship.org/uc/item/31c793kf.

Porayska-Pomsta, K. and Rajendran, G. (2019). Accountability in human and Artificial Intelligence decision-making as the basis for diversity and educational inclusion. In Knox, J., Wang, Y., and Gallagher, M., editors, *Artificial Intelligence and Inclusive Education: Speculative Futures and Emerging Practices*, pages 39–59. Springer Singapore, Singapore. Available at: https://doi.org/10.1007/978-981-13-8161-4_3.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526. Institute of Electrical and Electronics Engineers (IEEE). Available at: https://ieeexplore.ieee.org/document/4338389.

Preston, D. C. (2017). Untold barriers for black students in higher education: Placing race at the center of Developmental Education. Technical report, Southern Education Foundation, Atlanta, GA. Available at: https://eric.ed.gov/?id=ED585873.

Qian, L., Zhao, Y., and Cheng, Y. (2020). Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research*, 58(4):771–790. Available at: https://doi.org/10.1177/0735633119881472.

Ramesh, V., Parkavi, P., and Yasodha, P. (2011). Performance analysis of data mining techniques for placement chance prediction. *International Journal of Scientific & Engineering Research*, 2(8):1–7. Available at: https://www.researchgate.net/publication/254558569.

Roscoe, R. D., Crossley, S. A., Snow, E. L., Varner, L. K., and McNamara, D. S. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. *The 27th International Florida Artificial Intelligence Research Society Conference*, pages 393–398. Available at: https://files.eric.ed.gov/fulltext/ED585785.pdf.

Santos, R., Rodrigues, J., Branco, A., and Vaz, R. (2021). Neural text categorization with transformers for learning Portuguese as a second language. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings*, pages 715–726. Springer. Available at: https://link.springer.com/chapter/10.1007/978-3-030-86230-5_56.

Savary, A., Cordeiro, S., and Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. In Savary, A., Escartín, C. P., Bond, F., Mitrović, J., and Mititelu, V. B., editors, *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics. Available at: https://aclanthology.org/W19-5110/.

Schmitt, N. and Schmitt, D. (2020). *Vocabulary in Language Teaching*. Cambridge University Press. ISBN: 9781108476829.

Seßler, K., Fürstenberg, M., Bühler, B., and Kasneci, E. (2025). Can AI grade your essays? a comparative analysis of Large Language Models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472. Available at: https://dl.acm.org/doi/10.1145/3706468.3706527.

Singh, P. (2013). P-value, statistical significance and clinical significance. *Journal of Clinical and Preventive Cardiology*, 2(4):202–204. Available at: https://www.jcpcarchives.org/abstract/p-value-statistical-significance-and-clinical-significance-121.php.

Siyanova-Chanturia, A. and Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, 70(2):420–463. Available at: https://doi.org/10.1111/lang.12383.

Stavans, A. and Zadunaisky-Ehrlich, S. (2024). Text structure as an indicator of the writing development of descriptive text quality. *Journal of Writing Research*, 15(3):463–496. Available at: https://doi.org/10.17239/jowr-2024.15.03.02.

Suárez-Álvarez, J., Oliveri, M., Zenisky, A., and Sireci, S. (2023). Current assessment needs in adult education and workforce development: Summary report (assessment report no. 998). Technical report, Center for Educational Assessment, University of Massachusetts Amherst. Available at: https://createadultskills.org/system/files/ASAP%20Brief%201_Nov%202023_Needs%20Assessment.pdf.

Temnikova, I., Jr., W. A. B., Hailu, N. D., Nikolova, I., Mcenery, T., Kilgarriff, A., Angelova, G., and Cohen, K. B. (2014). Sublanguage corpus analysis toolkit: a tool for assessing the representativeness and sublanguage characteristics of corpora. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language

Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/675_Paper.pdf.

The College Board (2018). ACCUPLACER Program Manual. Retrieved from: https://www.aud.edu/media/nnmfzpv4/accuplacer-program-manual.pdf. The College Board, New York.

The College Board (2022). ACCUPLACER Program Manual. https://secure-media.collegeboard.org/digitalServices/pdf/accuplacer/accuplacer-program-manual.pdf. The College Board, New York.

U.S. Department of Education, Office of Educational Technology (2023). Artificial Intelligence and future of teaching and learning: Insights and recommendations. Technical report, U.S. Department of Education, Washington, DC. Available at: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf.

Wang, H., Dang, A., Wu, Z., and Mac, S. (2024a). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, 7:100326. Available at: https://doi.org/10.1016/j.caeai.2024.100326.

Wang, Z., Pang, Y., Lin, Y., and Zhu, X. (2024b). Adaptable and reliable text classification using Large Language Models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 67–74. Available at: https://ieeexplore.ieee.org/document/10917961.

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K. P., and François, T. (2022). FABRA: French aggregator-based readability assessment toolkit. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association. Available at: https://aclanthology.org/2022.lrec-1.130/.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge University Press. ISBN:9780521829243.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., and Fu, Q. (2025). Human-AI collaborative essay scoring: A dual-process framework with LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 293–305, New York, NY, USA. Association for Computing Machinery. Available at: https://doi.org/10.1145/3706468.3706507.

Xiao, F. (2010). Corpus creation. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, chapter 7, pages 147–166. Chapman & Hall/CRC Press, Boca Raton, FL, USA, 2 edition. eBook ISBN: 9780429149207.

Zachry Rutschow, E., Edgecombe, N., and Bickerstaff, S. (2021). A brief history of Developmental Education reform. Available at: https://postsecondaryreadiness.org/research/history-developmental-education-reform/.

Zhang, K. and Aslan, A. B. (2021). AI Technologies for Education: Recent Research & Future Directions. *Computers and Education: Artificial Intelligence*, 2:100025. Available at: https://www.sciencedirect.com/science/article/pii/S2666920X21000199.

Appendix

A (DevEd Level) Sample Essay Responses from Corpus with ACCUPLACER Automatic Classification Scores

Below are two sample essay responses produced by college-intending students who took the ACCUPLACER test during the period of this investigation. The classification levels assigned by ACCUPLACER (DevEd Level 1 and DevEd Level 2) and the rationales provided by the system appear after each text.

DevEd Level 1 Sample Essay Response

Text ID: 40

No good results come when you tell the truth. And bad results come when you tell a story. But particulary i have a friend she tell's me story's all the time and im starting to think she not a friend. Because everytime we doing fun things she come out the bluesky she will say something like your not a real friend . And i stop her when she telling me made up things and tell her nothing good gone come to you if you keep telling me story's that not even true. Just so she can make her self feel good and thats what i get for thinking i had a friend. but the whole time she was lien so i backed up off our friendship because i been done wrong by friends. But i also been good to my old friends and that played a big part of me taking peoples serious and belive peoples word. And i also started to changed up the people i was hanging with because i wanted the truth from the people like i gave them .and sometime you have to let them all go so you good things start to happen to those who wait and tell the truth.

Test Results

Test Name	Score
TCCWritePlacer	1

Holistic Score Description
The essay demonstrates no mastery of on-demand essay writing.

Dimension Title	Dimension Description
Purpose and Focus	Your response lacks clear purpose and focus.
Organization and Structure	Your response demonstrates poor organization of ideas.
Development and Support	Your response needs additional ideas and support.
Sentence Variety and Style	Your response shows limited ability to vary sentence length and apply appropriate vocabulary.
Mechanical Conventions	Your response shows poor control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking	Your response shows insufficient reasoning and lacks complexity of thought.

Course Placements

Writing Foundations I

You may enroll in ENGL 0923 Writing Foundations I.

DevEd Level 2 Sample Essay Response

Text ID: 138

Learning history gives you a viewpoint of the mistakes and triumphs from the past. I believe by erasing history will be detrementrial to future generations due to having different viewpoints and not knowing the actual facts of the past. History teaches the cause and effects of past periods, such as civil rights war, slavery, holocaust, WWI, and WWII. History teaches us the experience of past generations and how they overcame their adversities. I feel current generations knowing history is preventing certain ethnicities from experiencing further proverty and discrimination. Not learning about history does not erase the past, it prevents future generations from understanding how society operates and to continue improving human relations. Knowing and understanding history makes for improved behaviors for all.

Test Results

Test Name	Score
TCCWritePlacer	3

Holistic Score Description
The essay demonstrates little mastery of on-demand essay writing.

Dimension Title	Dimension Description
Purpose and Focus	Your response does not fully communicate purpose, and focus may be inconsistent.
Organization and Structure	Your response demonstrates limited organization of ideas.
Development and Support	Your response has limited support for your ideas.
Sentence Variety and Style	Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Mechanical Conventions	Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking	Your response shows limited clarity and complexity of thought.

Course Placements

(Summer 2025) Writing Foundations II

You may enroll in ENGL 0933 Writing Foundations II. Please see an advisor for course selection and verification of course placement.

B Annotation Framework for Developmental Writing

Two expert linguists developed the following annotation framework to capture a range of orthographic, grammatical, lexical, semantic, and discursive features found in the writing of students enrolled in Developmental Education (DevEd). The framework encompasses 21 distinct features, systematically grouped into four major categories of textual patterns, each accompanied by definitions and corresponding tags.

To support interpretability and pedagogical application, each feature is illustrated through carefully curated tagged examples drawn from the corpus. These features are additionally marked for polarity: (–) to indicate deviations from expected language proficiency, and (+) to signal evidence of advanced or proficient language usage.

1. **Orthographic patterns:** patterns including features such as punctuation, spelling, and contractions that often indicate early-stage writing development. For annotation purposes, the grapheme-level features (*) were grouped and considered as a single feature <ORT>.

Orthographic Pattern	**Definitions**	**Tags**
(-) contractions	Lack of punctuation to form a contraction changing the meaning of a word, for example, the contraction of a pronoun with a verb that with missing punctuation forms an adjective; missing apostrophe to indicate possession. Contractions are features not generally employed in formal academic writing.	<CONTRACT>
(-) grapheme_addition*	Addition of letter(s) or syllable(s), which alters the correct spelling of a word.	<ADD>
(-) grapheme_omission*	Omission of letter(s) or syllable(s), which alters the correct spelling of a word.	<OMIT>
(-) grapheme_transposition*	Inverting or transposing of letter(s) or syllable(s), which alters the correct spelling of a word.	<TRANSPOSE>
(-) grapheme_capitalization*	Misuse of capital letters in a sentence.	<CAPS>
(-) punctuation_used	Misuse of punctuation, for example, to join or separate clauses. Excessive use of the same punctuation mark within a large string, which reduces readability if used to split the text into shorter sentences.	<PUNCT>
(-) word_boundary_split	Division of words that are/must be joined as one graphical unit due to the origin of word/etymology or function.	<WORDSPLIT>
(-) word_boundary_merged	Joined word as one unit that is/must be divided into two due to origin of word/etymology or function. Single token used instead of compound nouns.	<WORDMERGED>

Continued on next page.

Continued from previous page.

Orthographic Patterns	Tagged Examples
contractions	<CONTRACT:it's>its</CONTRACT> <CONTRACT:today's>todays</CONTRACT>
grapheme_addition	<ORT:chosen>choosen</ORT>
grapheme_omission	<ORT:miserable>misrable</ORT> <ORT:pregnant>pregangt</ORT>
grapheme_transposition	<ORT:almost>alomst</ORT> <ORT:conclusion>conculsion</ORT>
grapheme_capitalization	<ORT:United States of America>united states of america</ORT>
punctuation_used	I need to protect<PUNCT:myself. I>myself, I</PUNCT> also need help. Stronger than ever <PUNCT:before. I>before, I</PUNCT> had no money.
word_boundary_split	Noun: <WORDSPLIT:somebody>some body</WORDSPLIT> Adjective: <WORDSPLIT:everyday>every day</WORDSPLIT> Adverb: <WORDSPLIT:sometimes>some times</WORDSPLIT> Pronoun: <WORDSPLIT:whatever>what ever</WORDSPLIT>
word_boundary_merged	<WORDMERGED:food chain>foodchain</WORDMERGED> <WORDMERGED:real estate>realestate</WORDMERGED>

2. **Grammatical patterns:** patterns including grammatical features commonly targeted in learner corpora, including verb tense, agreement, and word omissions or additions.

Grammatical Patterns	Definitions	Tags
(-) pronoun_alternation_referential	Incorrect use of a pronoun that does not properly match its antecedent, breaking an anaphoric chain.	<ALTERN>
(-) verb_tense	Verb not conjugated in accordance with its respective tense; verb conjugation may exemplify a combination of two tenses to form one in particular.	<VTENSE>
(-) verb_agreement	Agreement between the subject and the conjugated form of the verb.	<VAGREE>
(-) verb_form	Verb is used in the infinitive instead of conjugated (in present progressive) or it appears conjugated with the preposition 'to,' which typically indicates its infinitive.	<VFORM>
(-) word_omitted	Part of speech omitted; left out, which interferes with the meaning of a statement, for example, a preposition.	<WORDOMIT>
(-) word_added	Insertion of [an unnecessary] word that does not add semantic value; example could be the insertion of a verb in front of a conjunction.	<WORDADD>
(-) word_repetition	Repetition of adverbs, nouns, among other parts of speech within the same sentence without adding semantic value.	<WORDREPT>

Continued on next page.

Continued from previous page.

Grammatical Patterns	Tagged Examples
pronoun_alternation_referential	Before <ALTERN:we>you</ALTERN> get home, let's stop by the bank.
verb_tense	I've <VTENSE:written>wrote</VTENSE>
verb_agreement	<VAGREE:we weren't>we wasn't</VAGREE>
verb_form	<VFORM:try not to emphasize>not try to emphasizing</VFORM>. [...] themselves <VFORM:when finding>to find</VFORM> a practical skill.
word_omitted	depends <WORDOMIT:on></WORDOMIT> how
word_added	passage <WORDADD>is</WORDADD> because
word_repetition	I truly do believe that our ability to change ourselves is <WORDREPT>truly</WORDREPT> unlimited.

3. **Lexical & semantic patterns:** patterns including features that introduce clarity to the meaning of the sentences in a text, such as word choice and use of connectives. Multiword expressions (MWE) are also included in this category. For annotation purposes, in the initial stages of this study, all MWE (*) were considered as a single feature <MWE>.

Lexical & Semantic Patterns	Definitions	Tags
(-) connectives	Linking words or phrases inadequately used at the beginning of a discourse (e.g., opening statement) or at its end.	<CONNECT>
(-) mischosen_preposition	Missing or incorrect grammatical use of a preposition.	<PREP>
(-) slang	Use of informal contractions and similar devices not employed in formal academic writing.	<SLANG>
(-) word_precision	Imprecise use of a word attending to its meaning within a sentence.	<PRECISION>
(+) compound_adjectives*	MWE functioning as an adjective (excluding copula verbs); includes prepositional phrases but excludes predicative nouns.	<ADJ>
(+) compound_adverbs*	MWEs functioning adverbially, modifying a verb, adjective, or other adverb.	<ADV>
(+) compound_conjunctions and_prepositions*	MWE functioning as either a conjunction (linked to a verb) or a preposition (linked to a noun).	<CONJ>, <PREP>
(+) compound_nouns*	Multi-word expressions (MWEs) functioning as a single noun unit.	<N>
(+) named_entities*	Proper names of specific people, organizations, places, etc.	<NE>
(+) phrasal_verbs*	Verb-particle constructions, which are semantically non-compositional and distributionally constrained.	<PHV>
(+) support_verb_constructions*	Support verb + predicative noun constructions that together function semantically like a simple verb (e.g., "make a decision").	<SVC>
(+) verbal_idioms*	Combinations of a verb with at least one argument (e.g., subject, object), which are distributionally frozen and/or semantically non-compositional.	<VIDIOM>

Continued on next page.

Continued from previous page.

Lexical & Semantic Patterns	Tagged Examples
connectives	<CONNECT:Let's>So, let's</CONNECT>
mischosen_preposition	He looked <PREP:up>on</PREP>
slang	<SLANG:going to>gonna</SLANG> <SLANG:got to>gotta</SLANG>
word_precision	<PRECISION:wish>desire</PRECISION> <PRECISION:space>room</PRECISION>
compound_adjectives	<MWE>straight and tall</MWE> <MWE>strong minded</MWE>
compound_adverbs	<MWE>by the way</MWE> <MWE>back in the old day</MWE>
compound_conjunctions_and_prepositions	Conjunction: <MWE>as long as</MWE> Preposition: <MWE>in spite of</MWE>
compound_nouns	<MWE>golden rule</MWE> <MWE>refresher course</MWE>
named_entities	Person: <MWE>Collin Powell</MWE> Org: <MWE>United Nations</MWE> Place: <MWE>Guatemala City</MWE>
phrasal_verbs	<MWE>fit in</MWE> <MWE>mess up</MWE>
support_verb_constructions	<MWE>have no clue</MWE> <MWE>make mistakes</MWE>
verbal_idioms	<MWE>see the bigger picture</MWE> <MWE>weigh the pros and cons</MWE>

4. **Discursive patterns:** patterns including features reflecting a student's ability to structure extended arguments, express abstract thinking, and engage in academic-level discourse, which are skills essential for success in college-level writing.

Discursive Patterns	Definitions	Tags
(+) argumentation_with_reason	Argumentation of key concepts; supports one's position with reasoning and evidence of causal relation.	<REASON>
(+) argumentation_with_examples	Argumentation of key concepts; supports one's position with relevant examples.	<EXAMPLE>
(-) fictional_we	Fictional or imaginary representation of a person by the pronoun *we*.	<FICTIONAL_WE>
(-) fictional_you	Fictional or imaginary representation of a person by the pronoun *you*.	<FICTIONAL_YOU>

Continued on next page.

Continued from previous page.

Discursive Patterns	**Tagged Examples**
argumentation_with_reasoning	Evidence of words such as: *because*; *since*; *due to*; *owing to*. Coordinated instances count too.
argumentation_with_examples	Evidence of words that introduce examples: *as proof*; *to give an idea*; *for example*; *to illustrate*; *to show what I mean*; *e.g.*; *i.e.*
fictional_you	<FICTIONAL_YOU>You</FICTIONAL_YOU> really do not want to get out of bed, but <FICTIONAL_YOU>you</FICTIONAL_YOU> decide <FICTIONAL_YOU>you</FICTIONAL_YOU> have to [...].
fictional_we	<FICTIONAL_WE>we</FICTIONAL_WE> have the control over <FICTIONAL_WE>our</FICTIONAL_WE> happiness.

Annotation Guidelines for Evaluating Writing Proficiency in DevEd Texts Using ACCUPLACER Descriptors

The following pages present the annotation guidelines that were prepared by two expert linguists for a DevEd text classification task focused on ACCUPLACER's six textual criteria: Mechanical Conventions, Sentence Variety & Style, Development & Support, Organization & Structure, Purpose & Focus, and Critical Thinking. Due to the limited definitions available in the literature produced by ACCUPLACER, some definitions have been enhanced (Mechanical Conventions and Organization & Structure) with more precise assessment criteria (e.g., numerical scale to assess typos). This document has been referred to in (Da Corte and Baptista (2024b)).

Continued on next page.

Guidelines to Participate in a Classification Task Assessing Writing Proficiency in DevEd Courses

Purpose of research:

This research concentrates on the writing patterns of students who speak English as their first language. It aims to explore their knowledge of their first language and their comfort level using written English for communication in college settings. The objective is to identify common patterns students exhibit throughout their college careers and how these patterns may sometimes hinder their successful participation in and completion of academic programs. The goal is to uncover these patterns and develop linguistic resources that facilitate access to educational opportunities and aid in enhancing the writing skills of these students.

Instructions to raters:

Based on your qualifications, you have been selected to participate in a classification task involving two main steps: *(i) assess each essay based on 6 textual criteria*; and (ii) *produce an overall classification by skill level corresponding to DevEd Level 1, 2, or College level (demonstrating no need for DevEd)*.

As a rater in this study:

You will read and evaluate 25, randomly selected, short essays using the following the following 6 textual criteria:

- Mechanical Conventions
- Sentence Variety & Style
- Development & Support
- Organization & Structure
- Purpose & Focus
- Critical Thinking

A secured Google Forms document has been created to house the essays you will read and assess. Each essay will appear on an individual page. After assessing each essay, you will provide an overall assessment of the text and classify it according to 3 levels:

- Level 1_DevEd
- Level 2_DevEd
- Level 3_College-level

Each of the essay samples provides the respective definitions of the 6 textual criteria and the 3 classification levels above. To support you in this process, a document with some assessment examples is provided. Please carefully review the document before beginning your assessment, as it provides two examples—one for Level 1_DevEd and one for Level 2_DevEd—on how to navigate and complete this annotation task.

Please note that college students wrote the texts at the onset of their college careers. They are presented in their original form and assigned a unique identifier to anonymize the authors. The risks associated with this questionnaire are minimal. Some essays may cause some discomfort, particularly if the authors have experienced hardships or other life events with which you relate.

Your participation in this task is voluntary, and you may withdraw from it at any time. Thank you for taking the time to support this study and the overall access to educational opportunities.

Assessment Example # 1

Topic/prompt: Unlimited Change

Student A

The ability to change oneself is truly unlimited. 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. I'm in the career field that I not only enjoy but that I'm truly passionate about. I didn't come from a well off family. My mother was a young pregangt teen who had more kids then she had money to take care of.
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 who was a young mother depending on others to help take care of her children. Change is going to happen no matter what we choose to do, how we react to that change and if we chose to let it control us or if we decide to control it is what it truly boils down to. I've choosen to not let anything control my life but me. So every change, every desire that I wanted to happen or to accive I let nothing get in my way.
Determination is the key factor in how unlimited your change is. I believe that it depends how a few key factors; one of those factors being determination, another being how badly do you want it. If you truly want it bad enough then you will face all oods and defie them. You will allow nothing to stand in you way even if one of those objects in your way is yourself. You've got to ask yourself how badly do you want this change in ourselves or within our lifers.
In conculsion I truly do believe that our ability to change ourselves is truly unlimited if only we have the will power, determination and the drive to go after the things that we want changed.

Analysis:

1. **Mechanical conventions** refer to the extent to which the text expresses ideas using standard English. As you rate the text within this category, consider the following elements: spelling, grammar, and punctuation. In terms of **mechanical conventions**, the text is:

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

Several typos were noticed.

- 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.

Run-on sentences are evident throughout the text, 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.*

Punctuation signs are either not present or used incorrectly throughout the text, 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.*

Contractions, e.g., *didn't; I'm;* or the use of slang, e.g., *gonna;* appear on the text. These linguistic features are not used in common academic writing.

2. **Sentence variety & style** refer to the extent to which the text demonstrates control of vocabulary, voice, and structure. As you rate the text within this category, consider the following elements: sentence length, sentence structure, usage, tone, vocabulary, and voice. In terms of **sentence variety and style**, the text is:

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

Several sentences (or parts of them) repeat the same information at least twice, e.g., *change, truly*.

Sentence length, at times, exceeds 15 words, which is the average standard for improved readability and comprehension, e.g., *In conculsion I truly do believe that our ability to change ourselves is truly unlimited if only we have the will power, determination and the drive to go after the things that we want changed.*

The text exemplifies some word choices that are not clear or applicable to the context, e.g., the use of *desire* instead of *wish* or *dream* in *every desire that I wanted to happen*.

3. **Development & support** refer to the extent to which the text's ideas are developed and supported. As you rate the text within this category, consider the following elements: point of view, coherent arguments, evidence, and elaboration. In terms of **development and support**, the text is:

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

There is little evidence of words that introduce examples: *as proof; to give an idea; for example; to illustrate; to show what I mean; e.g.; i.e*. or evidence of words such as *because*, *since*, *due to*; *owing to*.

In the following sentence: *I didn't come from a well off family. My mother was a young pregangt teen who had more kids then she had money to take care of,* conjunctions such as *since* or *due to* could have enhanced the development of the sentence, for example: *I did not come from a well-off*

family since *my mother was a young pregnant teen who had more kids than she could afford.*

4. **Organization & structure** refer to the extent to which the text is ordered and connected. As you rate the text within this category, consider the following elements: introduction, thesis, body paragraphs, transitions, conclusions. In terms of **organization and structure**, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

The text was reasonably structured. There are four distinct paragraphs: an introduction with a semi-developed thesis and two supporting points: *family* and *determination*; the next two paragraphs attempt to provide details about these supporting points; and the text ends with a conclusion, connected with the original prompt.

5. **Purpose & focus** refer to the extent the text presents information in a unified and coherent manner, clearly addressing an issue. As you rate the text, consider the following elements: unity, consistency, coherence, relevance, and audience. In terms of **purpose and focus**, the text is:

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

Although the content of the text is relevant to the suggested topic and applies to the intended audience, it is written at a level that does not consistently (and accurately) showcases the use of proper English.

There is evidence of some informal language used (as in casual conversations) that may not align with the requirements of academic writing, e.g., *he constantly told me how I was gonna end up just like my mother* [...]

An idea is communicated, but not in a very coherent way, e.g., *Determination is the key factor in how unlimited your change is. I believe that it depends how a few key factors; one of those factors being determination.*

6. **Critical thinking** refers to the extent to which the text communicates a point of view and demonstrates reasoned relationships among ideas. As you rate the text, consider the following elements: accuracy, fairness, breadth, relevance, clarity, depth, precision, and logic. In terms of **critical thinking**, the text is:

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

Words are not used accurately with the intended meaning as presented in previous examples.

Only one point of view is presented when it comes to explaining *unlimited change*; no counterarguments are raised.

Some *colloqualisms* (words that are not used in formal communication) and *idioms* (expressions used not in their literal meaning) are evident in the text, e.g., *You will allow nothing to stand in you way; You've got to ask yourself.*

Repetition of words; two (or more) words are used to describe one idea, e.g., *if only we have the will power, determination and the drive to go after the things that we want changed.*

The texts presents a variety of positions but supporting evidence to each of the arguments is not provided. Reasoning is not sound.

Based on your assessment of the writing sample recently reviewed, what level would you assign to it:

- **Level 1_DevEd:** texts in this level indicate that development is needed in the overall use of the English language. Several of these skills need to be developed: spelling, punctuation, grammar, and sentence and paragraph structure.

- Level 2_DevEd: texts in this level indicate that support in specific areas related to essay writing is needed. Some of these areas include sentence structure, punctuation, editing, and revising.

- Level 3_College-level: texts in this level are written accurately and showcase the use of proper English at the college/academic level. These texts successfully communicate and support a very specific idea or point of view.

Assessment Example # 2

Topic/prompt: Success

Student B

Success in life is earned because as Colin Powell said "There are no secrets to success". It matters how much work youre going to put because stuff in life just doesnt happen if you dont try. So put it in example like this you are studying for a test and you do not study at all you just wait until the last day to study, rather then you have been studying you have been preparing for the test you are more up to be more successful then the one waitng until the last second. But with having success somtimes you fail and it is okay its life thing happen, maybe it was not your time yet or maybe you set yourself for a bigger chance for something better. When Colin Powell also stated "Success is the result of preparation, hardwork, and learning from failure". It shows me of what i have seen like my friend was practicing on basketball working hard and he would put in the work and when he tried out he didnt make it he was sad and i could tell he really wanted to play on the team but he did not give up he tried and tried, then basketball team had second tryouts and he made the team. That goes to show that even if you fail you learn from it and you keep going you keep trying and never give up on the road to success. I also look at Karl Popper and it states "and in all fields of life there have always been of great merit who did not succeed". I agree but i disagree because alot of people have failed but are still trying they did not give up they are not wasting time they have to earn it like and maybe it was not the right time for them but they are not giving up you learn from your failure and you keep on going.

Analysis:

1. **Mechanical conventions**

0 – deficient; **1 – below average**; 2 – above average; 3 – outstanding

1 - below average was chosen because:

Several typos were noticed.

- 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.

Run-on sentences are evident throughout the text, e.g., *It shows me of what i have seen like my friend was practicing on basketball working hard and he would put in the work and when he tried out he didnt make it he was sad and i could tell he really wanted to play on the team but he did not give up he tried and tried.*

Punctuation signs are either not present or used incorrectly throughout the text, e.g., *I agree [,] but i disagree because alot of people have failed but are still trying[.] they did not give up[;] they are not wasting time[.] they have to earn it like and maybe it was not the right time for them [,] but they are not giving up[.] you learn from your failure and you keep on going.*

Contractions, e.g., *It matters how much work youre going to put because stuff in life just doesnt happen if you dont try;* or the use of slang, e.g., *stuff in life ;* appear on the text. These linguistic features are not used in common academic writing.

Capitalization of proper nouns or subject pronouns is not consisten: *It shows me of what i have seen; I agree but i disagree because alot of people have failed.*

2. **Sentence variety & style** refer to the extent to which the text demonstrates control of vocabulary, voice, and structure. As you rate the text within this category, consider the following elements: sentence length, sentence structure, usage, tone, vocabulary, and voice. In terms of **sentence variety and style**, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

Although certain words repeat often, e.g., *keep going; you keep trying; you keep on going; did not give up,* repetition of sentences is not evident.

Sentence length, at times, exceeds 15 words, which is the average standard for improved readability and comprehension, e.g., *But with having success somtimes you fail and it is okay its life thing happen, maybe it was not your time yet or maybe you set yourself for a bigger chance for something better.*

The text evidences a fair choice and use of the words to articulate ideas, e.g., *That goes to show that even if you fail you learn from it and you keep going you keep trying and never give up on the road to success*.

3. **Development & support** refer to the extent to which the text's ideas are developed and supported. As you rate the text within this category, consider the following elements: point of view, coherent arguments, evidence, and elaboration. In terms of **development and support**, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

There is evidence of words that introduce examples: *as Colin Powell said; I also look at Karl Popper and it states; So put it in example like this*.

The text includes a personal discussion of the evidence accompanying every quote, e.g., *I also look at Karl Popper and it states "and in all fields of life there have always been of great merit who did not succeed". I agree but i disagree because alot of people have failed but are still trying they did not give up*.

The text provides reasoning and evidence (such as anecdotal or textual) as evidenced in the example provided.

4. **Organization & structure** refer to the extent to which the text is ordered and connected. As you rate the text within this category, consider the following elements: introduction, thesis, body paragraphs, transitions, conclusions. In terms of **organization and structure**, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

The text was reasonably structured. Although the entire text is presented as one, long paragraph, the ideas presented can be easily separated. There is a thesis statement developed and three supporting points: *secrets to success*, *the result of planning and preparing for a task*, *results of no success*. Within the text details and examples are provided to supporting the points presented; the text ends with a semi-developed conclusion, connected with the original prompt.

5. **Purpose & focus** refer to the extent the text presents information in a unified and coherent manner, clearly addressing an issue. As you rate the text, consider the following elements: unity, consistency, coherence, relevance, and audience. In terms of purpose and focus, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

The content of the text is relevant to the suggested topic and applies to the intended audience. It is written at a level that, with some editing and revising will more consistently showcase the use of proper English.

The text matches the main points as they are previewed in the thesis statement.

There is some evidence of informal language being used, but this does not interfere with or distract one from reading the entire text.

Ideas throughout the text are communicated in a sound way.

6. **Critical thinking** refers to the extent to which the text communicates a point of view and demonstrates reasoned relationships among ideas. As you rate the text, consider the following elements: accuracy, fairness, breadth, relevance, clarity, depth, precision, and logic. In terms of critical thinking, the text is:

0 – deficient; 1 – below average; **2 – above average**; 3 – outstanding

2 - above average was chosen because:

The text evidences that words are used adequately with the intended meaning.

Although the text presents a point of view, a counterargument is provided, e.g., *I agree but i disagree because alot of people have failed but are still trying*. The text provides some convincing arguments about the main claim or topic: *success*.

Several *colloqualisms* (words that are not used in formal communication) and *idioms* (expressions used not in their literal meaning) are present in the text, e.g., *stuff in life; it was not your time; keep going; wasting time; give up; not the right time.*

Some repetition of words is observed in the description of ideas, as previously exemplified in other categories.

The texts presents different positions and the respective supporting evidence to each of the arguments made. Reasoning is sound.

Based on your assessment of the writing sample recently reviewed, what level would you assign to it:

- **Level 2_DevEd:** texts in this level indicate that support in specific areas related to essay writing is needed. Some of these areas include sentence structure, punctuation, editing, and revising.

- Level 1_DevEd: texts in this level indicate that development is needed in the overall use of the English language. Several of these skills need to be developed: spelling, punctuation, grammar, and sentence and paragraph structure.

- Level 3_College-level: texts in this level are written accurately and showcase the use of proper English at the college/academic level. These texts successfully communicate and support a very specific idea or point of view.

C Multiword Expressions (MWE) Examples from Native English Speakers' DevEd-Text Productions

This table presents 25 examples of MWE detected in the written production of native English speakers within a DevEd context. MWE categories: Phrasal Verbs (PHV); Compound Nouns (N); Compound Adverbs (ADV); Verbal Idioms (VIDIOM).

MWE List
fit_in $$ PHV
focus_on $$ PHV
forced_into $$ PHV
gave_up $$ PHV
get_back $$ PHV
get_out_of_something $$ PHV
getting_myself_into $$ PHV
give_in $$ PHV
full_control $$ N
full_time_job $$ N
golden_rule $$ N
hard_way $$ N
hard_work $$ N
high_school $$ N
higher_education $$ N
pretty_much $$ ADV
some_ways $$ ADV
soon_to_be $$ ADV
sooner_or_later $$ ADV
step_by_step $$ ADV
time_and_time_again $$ ADV
bump_in_the_road $$ VIDIOM
busted_her_rear $$ VIDIOM
catches_their_eye $$ VIDIOM
get_lucky $$ VIDIOM

Multiword expressions (MWE) from Non-native and Native Spanish Speakers' Text Productions

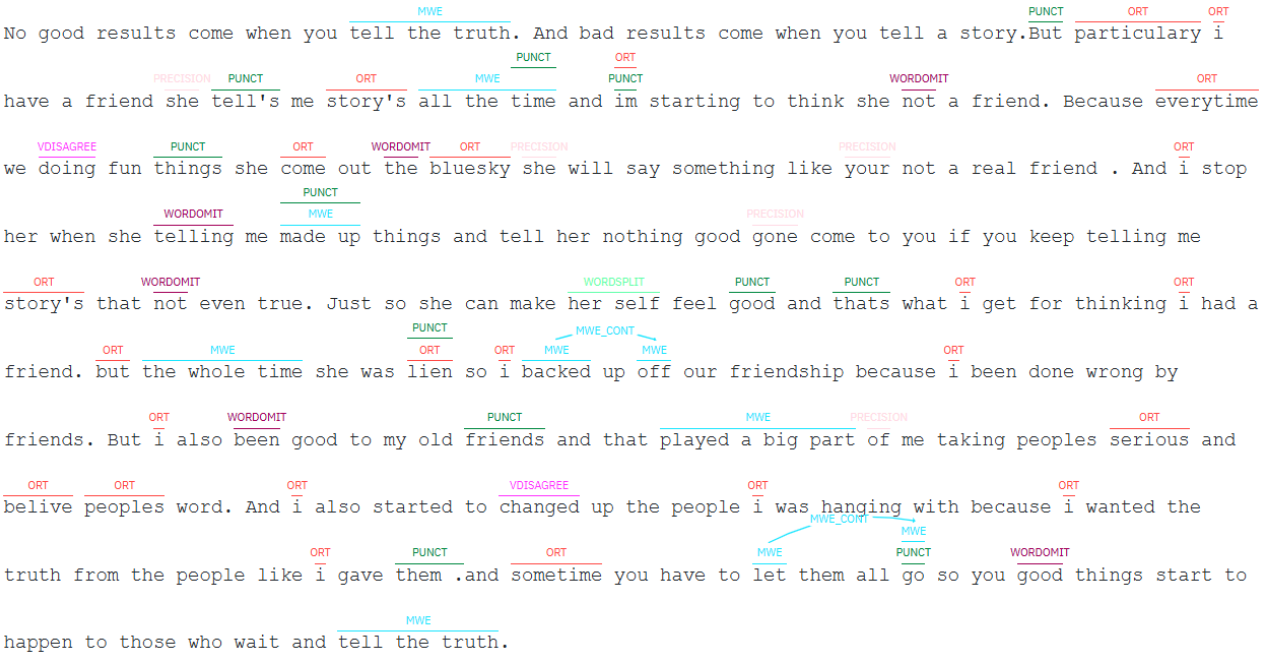
This table presents 50 examples of multiword expressions (MWE) detected in the written production of non-native and native Spanish speakers in a Grammar and Composition course. The list includes 25 MWEs from each group, placed side by side for comparison. MWE categories: Compound Adverbs (ADV); Compound Nouns (N); Compound Prepositions (PREP); Support Verb Constructions (SVC); Verbal Idioms (VIDIOM).

MWE from non-native speakers' texts	MWE from native speakers' texts
A_veces $$ ADV	amor_de_padre $$ N
alma_gemela $$ N	asumir_##_compromiso $$ SVC
alrededor_del_mundo $$ ADV	base_militar $$ N
centro_comercial $$ N	cometiendo_infidelidad $$ SVC
citas_a_ciegas $$ N	clubes_de_baile $$ N
Con_el_pasar_del_tiempo $$ ADV	cuentan_el_uno_con_el_otro $$ VIDIOM
dar_gracias $$ SVC	De_vez_en_cuando $$ ADV
Desde_mi_perspectiva $$ ADV	en_lo_que_respecta_a $$ PREP
dieron_las_herramientas $$ VIDIOM	En_lugar_de $$ PREP
directores_de_cine $$ N	experiencia_personal $$ N
dispositivo_inteligente $$ N	Fayetteville_Arkansas $$ NE
En_algunas_ocasiones $$ ADV	formas_de_entretenimiento $$ N
estar_al_tanto $$ VIDIOM	hacer_diligencias $$ SVC
futuras_generaciones $$ N	hermana_menor $$ N
Hace_##_##_tiempo $$ ADV	interponen_en_el_camino $$ VIDIOM
invitar_a_salir $$ VIDIOM	ir_de_compras $$ VIDIOM
irse_de_la_casa_de_sus_padres $$ VIDIOM	momentos_de_necesidad $$ N
juega_un_papel $$ SVC	mostrar_respeto $$ SVC
libertad_de_prensa $$ N	ofrece_paz $$ SVC
medios_sociales $$ N	Oklahoma_City $$ NE
mirar_hacia_un_futuro $$ VIDIOM	palabras_sabias $$ N
nos_falta_camino_por_recorrer $$ VIDIOM	ponerse_pesado $$ VIDIOM
padres_de_familia $$ N	Por_ejemplo $$ ADV
pista_de_hielo $$ N	ritmo_de_vida $$ N
poner_etiquetas $$ SVC	ser_queridos $$ N

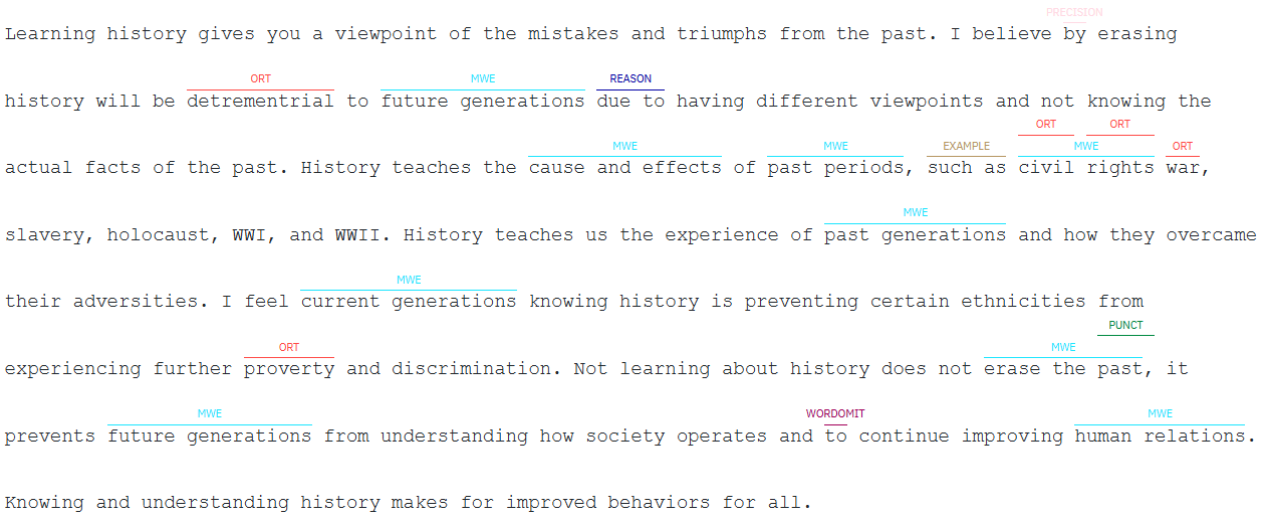
D DevEd and College Level Text Samples Annotated

Below are two sample essays from the corpus assembled for this investigation. Both texts were annotated through LABELBOX, with the annotations shown reflecting the final agreed-upon labels. The first essay represents a DevEd Level 1 text, while the second corresponds to a DevEd Level 2 text.

DevEd Level 1 Annotated Sample Essay Response; Text ID: 40



DevEd Level 2 Annotated Sample Essay Response; Text ID: 138



E Gold Standard for DevEd Writing

Gold standard for the 6 linguistic descriptors and skill levels with a corpus of 100 DevEd text samples.
Data presentation:
ID = document ID; **MC** = Mechanical Conventions; **SVS** = Sentence Variety & Style; **DS** = Development & Support; **OS** = Organization & Structure; **PF** = Purpose & Focus; **CT** = Critical Thinking; and **DevEd** = Development Education Level ("1" or "2").
Likert scale: deficient = **0**; below average = **1**; above average = **2**; and outstanding = **3** = outstanding.

1,0,1,1,1,1,1,1
3,0,1,1,1,1,1,1
4,1,1,0,0,1,1,1
5,0,1,2,1,2,2,1
9,1,1,1,1,2,1,1
12,1,1,1,2,1,1,1
13,2,1,1,1,2,2,1
21,1,2,1,1,2,1,1
23A,0,0,1,0,1,0,1
23B,0,1,2,1,2,1,1
24A,2,2,2,2,2,1,2
24B,1,1,1,0,1,1,1
25,0,1,1,1,1,0,1
26,0,1,1,1,1,1,1
27,1,1,0,0,0,0,1
29A,0,1,1,0,1,0,1
29B,0,1,1,1,1,1,2
38,1,2,2,1,2,2,1
40A,0,1,2,1,2,2,1
40B,0,0,1,0,0,1,1
40C,1,1,1,1,2,2,1
45,2,2,0,1,1,1,1
48,0,1,1,0,1,1,1
49A,1,1,2,2,2,2,2
49B,1,1,1,0,1,1,1
51,1,2,2,1,2,1,1
55,1,1,1,1,1,1,1
56A,1,1,0,0,0,0,1
56B,1,1,1,1,1,1,1
59,2,1,1,1,1,2,1
60,1,1,1,0,1,1,1
61A,1,1,2,2,1,2,2
61B,0,1,1,1,1,1,2
69A,0,1,2,1,1,1,1
69B,1,1,1,1,2,1,1
70,1,0,0,1,1,1,1
73A,1,1,1,1,1,1,2
73B,2,1,1,1,2,1,2
73C,1,1,1,1,1,1,1
74,0,1,1,2,1,1,1
76,0,0,1,1,1,1,1
77,0,1,2,2,2,2,2
78,0,0,1,0,0,1,1
80A,1,2,2,2,2,1,2
80B,1,0,0,1,1,1,1
81,0,1,1,0,1,1,1
82,2,1,1,1,1,1,2
83,1,1,1,0,1,1,1
85,1,1,2,2,2,2,2
90A,1,2,2,2,2,2,2
90B,0,1,1,1,1,1,1
92A,0,1,0,1,0,0,1
92B,1,1,1,0,0,0,1
93A,0,1,2,2,1,2,2
93B,1,1,1,1,1,1,1
94,2,1,1,1,1,1,1
95A,1,1,1,2,2,1,1
95B,0,0,0,0,0,1,1
96,1,1,1,1,1,1,1
99,1,1,1,1,1,1,1
100,2,2,2,2,3,3,2
103,0,1,2,1,1,1,1
104A,1,1,1,1,1,2,1
104B,2,2,1,1,2,2,2
104C,1,2,1,1,1,1,2
106,1,1,2,2,2,2,1
110,1,2,1,1,1,1,1
113,0,2,1,0,1,1,1
115,0,1,1,0,1,1,1
116,1,1,1,1,1,1,1
118,1,1,1,1,1,1,1
119,2,3,2,2,3,2,2
120,2,1,0,1,2,2,1
123,3,2,1,1,1,1,1
124,1,2,1,1,2,1,1
125A,2,2,3,2,3,2,2
125B,1,1,1,1,1,2,1
125C,1,1,1,1,1,0,1
126,1,1,1,1,2,1,1
132,1,1,2,2,1,2,1
134,2,2,3,3,2,3,2
135,1,1,0,1,1,0,1
138,1,1,1,1,1,1,1
140,3,2,1,3,2,1,2
143A,2,2,3,2,2,2,2
143B,1,2,2,2,2,1,1
147,2,2,1,1,1,2,2
150,0,2,1,1,1,2,1
151A,1,2,1,1,2,1,1
151B,1,1,2,1,2,2,1
152,0,1,1,0,1,1,1
163,2,1,1,1,2,2,1
174,2,2,2,2,2,3,2
178,1,2,2,2,2,2,2
180,2,2,2,2,2,2,2
184,2,2,3,2,2,2,2
187A,1,2,2,2,2,2,1
193,2,3,2,2,3,2,2
198,2,1,1,2,2,1,2
199,3,2,2,1,3,1,1

F Summary of Top 31 Features by Information Gain Scores

The table below summarizes the top 31 features identified through the Information Gain method in (Da Corte and Baptista (2025c)). These features were drawn from three sources: Coh-Metrix, CTAP, and DES. These top feature reflect a mix of lexical, syntactic, and orthographic dimensions of writing proficiency that were most predictive of placement outcomes in supervised ML experiments conducted as part of this investigation.

Rank	Source	Feature	Info. Gain
1	Coh-Metrix	Paragraph length, number of sentences in a paragraph, mean	0.144
2	Coh-Metrix	Lexical diversity, VOCD, all words	0.142
3	CTAP	Lexical Sophistication: Easy noun tokens (NGSL)	0.124
4	Coh-Metrix	LSA given/new, sentences, mean	0.124
5	CTAP	Lexical Richness: Type Token Ratio (STTR NGSLeasy Nouns)	0.118
6	Coh-Metrix	Lexical diversity, MTLD, all words	0.115
7	CTAP	Number of POS Feature: Plural noun Types	0.109
8	Coh-Metrix	LSA given/new, sentences, standard deviation	0.109
9	**DES**	**EXAMPLE**	0.103
10	CTAP	POS Density Feature: Existential There	0.102
11	Coh-Metrix	Positive connectives incidence	0.100
12	Coh-Metrix	LSA overlap, adjacent sentences, mean	0.100
13	CTAP	Number of POS Feature: Existential there Types	0.099
14	CTAP	Number of POS Feature: Preposition Types	0.099
15	Coh-Metrix	WordNet verb overlap	0.098
16	CTAP	Number of Syntactic Constituents: Postnominal Noun Modifier	0.097
17	CTAP	Number of Word Types (including Punctuation and Numbers)	0.097
18	CTAP	Lexical Sophistication Feature: SUBTLEX Logarithmic Word Frequency (AW Type)	0.096
19	CTAP	Number of POS Feature: Existential there Tokens	0.096
20	CTAP	Lexical Sophistication: Easy noun types (NGSL)	0.094

Continued on next page

Continued from previous page

Rank	Source	Feature	Info. Gain
21	CTAP	Number of POS Feature: Verbs in past participle form Types	0.092
22	Coh-Metrix	LSA verb overlap	0.092
23	CTAP	Number of Unique Words	0.092
24	CTAP	Number of Tokens with More Than 2 Syllables	0.090
25	CTAP	Number of Word Types with More Than 2 Syllables	0.086
26	CTAP	Lexical Richness: HDD (excluding punctuation and numbers)	0.086
27	CTAP	POS Density Feature: Possessive Ending	0.086
28	CTAP	Number of Syntactic Constituents: Complex Noun Phrase	0.084
29	CTAP	Number of POS Feature: Plural noun Tokens	0.083
30	CTAP	Referential Cohesion: Global Lemma Overlap	0.083
31	**DES**	**ORT**	0.082

G (College-level) Sample Essay Response from Expanded Corpus with ACCUPLACER Automatic Classification Scores

College-level Sample Essay Response

Text ID: 269

The idea of obtaining money and material objects has become a plague in today's society. It has been said that, "greed will be the downfall of man". This ever-increasing lust has turned the idea of basic human care into a never ending paper chase. The proof is in the pudding the more you look around every day. Health insurance companies around the globe net billions of dollars from the pockets of the rich and the poor. Unforunately, the poor seem to suffer much more from this current system. We will go over a basic infrastructure of the healthcare system, how it divides working classes, and possible solutions for future generations. Insurance has become one of the most lucrative scams in the history of our nation. Many other countries have adopted this ideology as well. The idea that to receive proper medical, dental, and death benefits; that you must pay out large premiums. In return, you don't always receive the proper care or benefits that you paid for in the first place. Most people cannot even begin to think about affording the insurance in question. In my opinion, it's extremely sad to think about humans not receiving basic human care for their needs. Instead we are left to the current system to wonder and hope that we make it into our elderly years with a clean bill of health. The client pays a premium with many different stipulations attached. The higher the premium they pay, the "better" the benefits. Unfortunately, not everyone will obtain an upper-class status in life. This makes it incredibly difficult to obtain an insurance plan that can fulfill their needs. Not to mention; citizens with large families, the bill to foot for insurance is astronomically high. As a result, many lower income households and middle class households suffer from this current system. They have the choice of: pay rent and buy groceries or pay a ridiculously high premium in case something happens to them or a family member. The other option is to have no insurance at all, receive proper medical care, and then go into debt for many years paying off a hospital bill. There are many amazing solutions to this dystopian nightmare we have funneled ourselves into. Abolishing insurance companies and beginning a universal healthcare would be the most viable solution. A system where everyone chips in a little bit, usually in the form of a very small tax increase. Everyone at this point would receive the proper healthcare that they need. To add to this, universal basic income is another fascinating idea. The idea that all lower to middle income households will get a basic income to cover such necessities. Such as food, utility assistance, rent, and more. One day, we will all live in a utopian world where we learn that hoarding money and material possessions is not the definite goal for us as human beings. We will look back at this time in our history in embarrassment, as we look toward the stars for the advancement of the human race. There will be no corporations or systems set up to hurt the very people that built them. Working classes will be a thing of the past. The solutions for these complex issues will be set into motion and will work toward a brighter future. The question is, "when?". Well, that's up to you and I.

Test Results

Test Name	Score
TCCWritePlacer	5

Holistic Score Description
The essay demonstrates adequate mastery of on-demand essay writing.

Dimension Title	Dimension Description
Purpose and Focus	Your response does not fully communicate purpose, and focus may be inconsistent.
Organization and Structure	Your response demonstrates limited organization of ideas.
Development and Support	Your response has limited support for your ideas.
Sentence Variety and Style	Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Mechanical Conventions	Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking	Your response shows limited clarity and complexity of thought.

Course Placements - [redacted]

Composition I

You may enroll in ENGL 1113, Composition I, provided reading proficiency is met. Please see an advisor for course selection and verification of course placement.

College-level Annotated Sample Essay Response

An additional sample essay from the expanded corpus was included here. The text was classified by ACCUPLACER as college-level, and the system's feedback appeared immediately after the response. A human-annotated version of the same essay, produced through LABELBOX, is also provided for comparison. This description is positioned here, rather than at the beginning of this Appendix, due to formatting constraints.

Continued on next page.

The idea of obtaining money and material objects has become a plague in today's society. It has been said that, "greed will be the downfall of man". This ever-increasing lust has turned the idea of basic human care into a never ending paper chase. The proof is in the pudding the more you look around every day. Health insurance companies around the globe net billions of dollars from the pockets of the rich and the poor.

Unforunately, the poor seem to suffer much more from this current system. We will go over a basic infrastructure of the healthcare system, how it divides working classes, and possible solutions for future generations.

Insurance has become one of the most lucrative scams in the history of our nation. Many other countries have adopted this ideology as well. The idea that to receive proper medical, dental, and death benefits; that you must pay out large premiums. In return, you don't always receive the proper care or benefits that you paid for in the first place. Most people cannot even begin to think about affording the insurance in question. In my opinion, it's extremely sad to think about humans not receiving basic human care for their needs. Instead we are left to the current system to wonder and hope that we make it into our elderly years with a clean bill of health. The client pays a premium with many different stipulations attached. The higher the premium they pay, the "better" the benefits.

Unfortunately, not everyone will obtain an upper-class status in life. This makes it incredibly difficult to obtain an insurance plan that can fulfill their needs. Not to mention; citizens with large families, the bill to foot for insurance is astronomically high. As a result, many lower income households and middle class households suffer from this current system. They have the choice of: pay rent and buy groceries or pay a ridiculously high premium in case something happens to them or a family member. The other option is to have no insurance at all, receive proper medical care, and then go into debt for many years paying off a hospital bill.

There are many amazing solutions to this dystopian nightmare we have funneled ourselves into. Abolishing insurance companies and beginning a universal healthcare would be the most viable solution. A system where everyone chips in a little bit, usually in the form of a very small tax increase. Everyone at this point would receive the proper healthcare that they need. To add to this, universal basic income is another fascinating idea. The idea that all lower to middle income households will get a basic income to cover such necessities. Such as food, utility assistance, rent, and more.

One day, we will all live in a utopian world where we learn that hoarding money and material possessions is not the definite goal for us as human beings. We will look back at this time in our history in embarrassment, as we look toward the stars for the advancement of the human race. There will be no corporations or systems set up to hurt the very people that built them. Working classes will be a thing of the past. The solutions for these complex issues will be set into motion and will work toward a brighter future. The question is, "when?". Well, that's up to you and I.

H LLMs Prompts for DevEd and College Levels Writing Classification Task

This Appendix includes the full set of classification prompts used in the experiments described *Section 4*. It includes all four prompting conditions, Zero-shot, Few-shot, Enhanced, and Enhanced+, for both the 1-step and 2-step classification strategies. This document has been referred to in (Da Corte and Baptista (2025b)).

LLMs Classification Experiments Prompts: 1-step Classification

- **Zero-shot**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of three levels of writing:

  - DevEd Level 1
  - DevEd Level 2
  - College Level

  Classify and justify your decision based on these three levels.

- **Few-shot**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of three levels of writing:

  - **DevEd Level 1**: Texts with frequent grammar, spelling, and punctuation issues, and lacking cohesion;
  - **DevEd Level 2**: Texts with some structural improvements still needing targeted support; and
  - **College Level**: Texts demonstrating academic-level writing with minimal errors.

  Classify and justify your decision based on these three levels.

- **Enhanced**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of three strictly defined levels:

  - **DevEd Level 1**: Requires fundamental improvement in English usage. These texts often exhibit frequent foundational errors (e.g., orthographic mistakes, punctuation errors, inconsistent sentence structure). Cohesion and complexity tend to be lower, resulting in disorganized or unclear expression;
  - **DevEd Level 2**: Requires targeted support but shows improvement in writing fluency. Some lower-order issues remain, such as occasional punctuation or structural inconsistencies. These texts demonstrate increased presence of multiword expressions (MWE), more refined sentence construction, and emerging control over clarity and expression; and
  - **College Level**: Exhibits accurate English usage at an academic level. These texts display strong structural organization, refined grammatical control, and appropriate use of MWE, discursive strategies, and lexical precision. Errors are minimal and do not interfere with readability or comprehension.

  Also, provide a rationale referencing the following linguistic features:

- ORT (Orthographic errors): Spelling mistakes, capitalization issues.
- PUNCT (Punctuation errors): Missing or incorrect punctuation.
- GRAMM (Grammatical issues): Verb agreement, omitted words, incorrect phrasing.
- LEXSEM (Lexical and Semantic features): Precise vocabulary, use of multiword expressions (MWE).
- DISC (Discursive features): Argumentation, reasoning, example usage.

Classify and justify your decision based on these strict definitions.

- **Enhanced+**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of three strictly defined levels:

  - **DevEd Level 1:** Requires fundamental improvement in English usage. These texts often exhibit frequent foundational errors (e.g., orthographic mistakes, punctuation errors, inconsistent sentence structure). Cohesion and complexity tend to be lower, resulting in disorganized or unclear expression;
  - **DevEd Level 2:** Requires targeted support but shows improvement in writing fluency. Some lower-order issues remain, such as occasional punctuation or structural inconsistencies. These texts demonstrate increased presence of multiword expressions (MWE), more refined sentence construction, and emerging control over clarity and expression; and
  - **College Level:** Exhibits accurate English usage at an academic level. These texts display strong structural organization, refined grammatical control, and appropriate use of MWE, discursive strategies, and lexical precision. Errors are minimal and do not interfere with readability or comprehension.

  Evaluation criteria: Analyze for these descriptors:

  1. *Mechanical conventions* refer to the extent to which the text expresses ideas using standard English. As you rate the text within this category, consider the following elements: spelling, grammar, and punctuation.
  2. *Sentence variety and style* refer to the extent to which the text demonstrates control of vocabulary, voice, and structure. As you rate the text within this category, consider the following elements: sentence length, sentence structure, usage, tone, vocabulary, and voice.
  3. *Development and support* refer to the extent to which the text's ideas are developed and supported. As you rate the text within this category, consider the following elements: point of view, coherent arguments, evidence, and elaboration.
  4. *Organization and structure* refer to the extent to which the text is ordered and connected. As you rate the text within this category, consider the following elements: introduction, thesis, body paragraphs, transitions, conclusions.
  5. *Purpose and focus* refer to the extent the text presents information in a unified and coherent manner, clearly addressing an issue. As you rate the text, consider the following elements: unity, consistency, coherence, relevance, and audience.
  6. *Critical thinking* refers to the extent to which the text communicates a point of view and demonstrates reasoned relationships among ideas. As you rate the text, consider the following elements: accuracy, fairness, breadth, relevance, clarity, depth, precision, and logic.

  - **DevEd Level 1 Example:**

"*The ability to change oneself is truly unlimited. I agree with that statement because I know about changing myself [...].*"

1. In terms of mechanical conventions, the text is below average. Several typos were noticed.
2. In terms of *sentence variety and style*, the text is below average.
3. In terms of *development and support*, this text is below average.
4. In terms of *organization and structure*, this text is above average.
5. In terms of *purpose and focus*, this text is below average.
6. In terms of *critical thinking*, this text is below average.

– **DevEd Level 2 Example:**

"*Success in life is earned because as Colin Powell said 'There are no secrets to success' [...].*"

1. In terms of *mechanical conventions*, this text is below average.
2. In terms of *sentence variety and style*, the text is above average.
3. In terms of *development and support*, this text is above average.
4. In terms of *organization and structure*, this text is above average.
5. In terms of *purpose and focus*, this text is above average.
6. In terms of *critical thinking*, this text is above average.

– **College-level Example:**

"*Life can be very difficult, and even more so when we have to make choices [...].*"

1. In terms of *mechanical conventions*, this text is below average.
2. In terms of *sentence variety and style*, the text is above average.
3. In terms of *development and support*, this text is outstanding.
4. In terms of *organization and structure*, this text is outstanding.
5. In terms of *purpose and focus*, this text is above average.
6. In terms of *critical thinking*, this text is above average.

Classify and justify your decision based on these strict definitions.

LLMs Classification Experiments Prompts: 2-step Classification

- **Zero-shot**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of two levels of writing:

  - DevEd Level
  - College Level

  Classify and justify your decision based on these two levels.

- **Few-shot**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of two levels of writing:

  - **DevEd Level**: Texts with frequent grammar, spelling, and punctuation issues, and lacking cohesion. Texts may also need some structural improvements, still needing targeted support.
  - **College Level**: Texts demonstrating academic-level writing with minimal errors.

  Classify and justify your decision based on these two levels.

- **Enhanced**

  You are an expert in English writing assessment and Developmental Education. Classify the following text into one of two levels of writing:

  - **DevEd Level**: texts often contain frequent foundational errors such as orthographic mistakes, punctuation issues, and inconsistent sentence structure, leading to lower cohesion and unclear expression. Some texts still in this category, while exhibiting lower-order issues like occasional punctuation or structural inconsistencies, may show progress through an increased presence of multiword expressions (MWE), more refined sentence construction, and emerging control over clarity and expression.
  - **College Level**: texts exhibit accurate English usage at an academic level. These texts display strong structural organization, refined grammatical control, and appropriate use of MWE, discursive strategies, and lexical precision. Errors are minimal and do not interfere with readability or comprehension.

  Also, provide a rationale referencing the following linguistic features:

  - ORT (Orthographic errors): Spelling mistakes, capitalization issues.
  - PUNCT (Punctuation errors): Missing or incorrect punctuation.
  - GRAMM (Grammatical issues): Verb agreement, omitted words, incorrect phrasing.
  - LEXSEM (Lexical and Semantic features): Precise vocabulary, use of multiword expressions (MWE).
  - DISC (Discursive features): Argumentation, reasoning, example usage.

  Classify and justify your decision based on these strict definitions.

Enhanced+
You are an expert in English writing assessment and Developmental Education.
Classify the following text into one of two strictly defined levels:

- **DevEd Level**: texts often contain frequent foundational errors such as orthographic mistakes, punctuation issues, and inconsistent sentence structure, leading to lower cohesion and unclear expression. Some texts still in this category, while exhibiting lower-order issues like occasional punctuation or structural inconsistencies, may show progress through an increased presence of multiword expressions (MWE), more refined sentence construction, and emerging control over clarity and expression.
- **College Level**: texts exhibit accurate English usage at an academic level. These texts display strong structural organization, refined grammatical control, and appropriate use of MWE, discursive strategies, and lexical precision. Errors are minimal and do not interfere with readability or comprehension.

EVALUATION CRITERIA: Analyze for these descriptors:

1. *Mechanical conventions* refer to the extent to which the text expresses ideas using standard English. As you rate the text within this category, consider the following elements: spelling, grammar, and punctuation.
2. *Sentence variety and style* refer to the extent to which the text demonstrates control of vocabulary, voice, and structure. As you rate the text within this category, consider the following elements: sentence length, sentence structure, usage, tone, vocabulary, and voice.
3. *Development and support* refer to the extent to which the text's ideas are developed and supported. As you rate the text within this category, consider the following elements: point of view, coherent arguments, evidence, and elaboration.
4. *Organization and structure* refer to the extent to which the text is ordered and connected. As you rate the text within this category, consider the following elements: introduction, thesis, body paragraphs, transitions, conclusions.
5. *Purpose and focus* refer to the extent the text presents information in a unified and coherent manner, clearly addressing an issue. As you rate the text, consider the following elements: unity, consistency, coherence, relevance, and audience.
6. *Critical thinking* refers to the extent to which the text communicates a point of view and demonstrates reasoned relationships among ideas. As you rate the text, consider the following elements: accuracy, fairness, breadth, relevance, clarity, depth, precision, and logic.

- **DevEd-level Example:**

"*Success in life is earned because as Colin Powell said 'There are no secrets to success'*." [...]

1. In terms of *mechanical conventions*, this text is below average.
2. In terms of *sentence variety and style*, the text is above average.
3. In terms of *development and support*, this text is above average.
4. In terms of *organization and structure*, this text is above average.
5. In terms of *purpose and focus*, this text is above average.
6. In terms of *critical thinking*, this text is above average.

- **College-level Example:**

"*Life can be very difficult, and even more so when we have to make choices that don't have the intended outcome that we desired*. [...]"

1. In terms of *mechanical conventions*, this text is below average.

2. In terms of *sentence variety and style*, the text is above average.
3. In terms of *development and support*, this text is outstanding.
4. In terms of *organization and structure*, this text is outstanding.
5. In terms of *purpose and focus*, this text is above average.
6. In terms of *critical thinking*, this text is above average.

Classify and justify your decision based on these strict definitions.

I Comparison of Feedback by Gemini and Claude, following the Enhanced+ Prompting Strategy

Below are three sample essays from the 450-text corpus, assessed and classified by Gemini and Claude using the Enhanced+ prompting strategy. For context, each text also includes: (1) the gold standard classification; (2) the classification levels assigned by ACCUPLACER; and (3) the classification levels automatically assigned by both LLMs. Each essay is followed by the feedback provided by each LLM.

Text ID: 40

Classification Levels:
Gold Standard = DevEd Level 1; **ACCUPLACER** = DevEd Level 1; **Gemini** = DevEd Level 1; **Claude** = DevEd Level 1

Student Essay (ID 40)

No good results come when you tell the truth. And bad results come when you tell a story. But particulary i have a friend she tell's me story's all the time and im starting to think she not a friend. Because everytime we doing fun things she come out the bluesky she will say something like your not a real friend . And i stop her when she telling me made up things and tell her nothing good gone come to you if you keep telling me story's that not even true. Just so she can make her self feel good and thats what i get for thinking i had a friend. but the whole time she was lien so i backed up off our friendship because i been done wrong by friends. But i also been good to my old friends and that played a big part of me taking peoples serious and belive peoples word. And i also started to changed up the people i was hanging with because i wanted the truth from the people like i gave them .and sometime you have to let them all go so you good things start to happen to those who wait and tell the truth.

Model Rationales (Gemini vs. Claude)

Gemini Feedback:

The text contains frequent errors in spelling ("story's," "lien"), punctuation (missing commas, incorrect apostrophe usage), and grammar ("she tell's," "nothing good gone come"). Sentence structure is often inconsistent and lacks complexity, hindering clarity and cohesion. While the ideas are somewhat developed, the numerous foundational errors overshadow any potential for refined expression.

Claude Feedback:

This text exhibits characteristics of DevEd Level 1 writing. The rationale for this classification is as follows:

1. Mechanical conventions: The text shows frequent foundational errors in spelling (e.g., "particulary," "lien," "belive"), punctuation (missing apostrophes, inconsistent use of periods), and grammar (subject-verb agreement issues, incorrect verb tenses).

2. Sentence variety and style: The writing demonstrates low control over sentence structure, with run-on sentences and inconsistent capitalization. Vocabulary use is limited and repetitive.

3. Development and support: While the text attempts to convey personal experiences, the ideas are not well-developed or supported with clear examples.

4. Organization and structure: The text lacks a clear introduction, thesis, or conclusion. Ideas are presented in a disorganized manner without logical transitions between thoughts.

5. Purpose and focus: The writing struggles to maintain a consistent focus, jumping between different ideas without a clear, unified message.

6. Critical thinking: The text presents a simplistic view of truth-telling and friendship, lacking depth in analysis or consideration of alternative perspectives. Overall, this text exhibits the characteristics of DevEd Level 1 writing, with significant issues in foundational writing skills and a need for substantial improvement in clarity, organization, and expression of ideas.

Text ID: 138

Classification Levels:
Gold Standard = DevEd Level 2; **ACCUPLACER** = DevEd Level 2; **Gemini** = DevEd Level 2; **Claude** = DevEd Level 2

Student Essay (ID 138)

Learning history gives you a viewpoint of the mistakes and triumphs from the past. I believe by erasing history will be detrementrial to future generations due to having different viewpoints and not knowing the actual facts of the past. History teaches the cause and effects of past periods, such as civil rights war, slavery, holocaust, WWI, and WWII. History teaches us the experience of past generations and how they overcame their adversities. I feel current generations knowing history is preventing certain ethnicities from experiencing further proverty and discrimination. Not learning about history does not erase the past, it prevents future generations from understanding how society operates and to continue improving human relations. Knowing and understanding history makes for improved behaviors for all.

Model Rationales (Gemini vs. Claude)

Gemini Feedback:

The text exhibits foundational errors characteristic of DevEd writing. While the ideas are present, there are noticeable orthographic mistakes ("detrementrial," "proverty"), grammatical issues (inconsistent verb tense in the first sentence, missing articles), and punctuation problems (missing commas). The sentence structure, though generally understandable, lacks refinement, and these errors collectively affect the clarity and cohesion of the text. Although the text demonstrates some capacity for multiword expressions and expresses a fairly clear central idea, the cumulative impact of the mechanical and structural errors places it within the DevEd level.

Claude Feedback:

This text demonstrates characteristics of DevEd Level 2 writing, showing improvement in writing fluency but still requiring some targeted support. The text exhibits:

1. Mechanical conventions: Some minor errors (e.g., "detrementrial" instead of "detrimental", "proverty" instead of "poverty"), but overall better control than Level 1.
2. Sentence variety and style: Improved sentence structure and vocabulary use (e.g., "Learning history gives you a viewpoint of the mistakes and triumphs from the past"), showing more refined construction.
3. Development and support: The text presents a clear argument with supporting examples (e.g., mentioning specific historical events like civil rights war, slavery, holocaust, WWI, and WWII).
4. Organization and structure: The text has a logical flow, starting with a main idea and following with supporting points, though it lacks a formal introduction and conclusion.
5. Purpose and focus: The writing maintains a consistent focus on the importance of learning history, demonstrating improved coherence compared to Level 1.
6. Critical thinking: The text shows evidence of reasoned relationships between ideas (e.g., connecting historical knowledge to improved human relations and societal understanding). While there are still some areas for improvement, particularly in mechanical conventions, the overall quality of writing, including the presence of more complex ideas and better sentence construction.

Text ID: 269

Classification Levels:
Gold Standard = College-level; **ACCUPLACER** = College-level; **Gemini** = College-level; **Claude** = College-level

Continued on next page.

Student Essay (ID 269)

The idea of obtaining money and material objects has become a plague in today's society. It has been said that, "greed will be the downfall of man". This ever-increasing lust has turned the idea of basic human care into a never ending paper chase. The proof is in the pudding the more you look around every day. Health insurance companies around the globe net billions of dollars from the pockets of the rich and the poor. Unforunately, the poor seem to suffer much more from this current system. We will go over a basic infrastructure of the healthcare system, how it divides working classes, and possible solutions for future generations. Insurance has become one of the most lucrative scams in the history of our nation. Many other countries have adopted this ideology as well. The idea that to receive proper medical, dental, and death benefits; that you must pay out large premiums. In return, you don't always receive the proper care or benefits that you paid for in the first place. Most people cannot even begin to think about affording the insurance in question. In my opinion, it's extremely sad to think about humans not receiving basic human care for their needs. Instead we are left to the current system to wonder and hope that we make it into our elderly years with a clean bill of health. The client pays a premium with many different stipulations attached. The higher the premium they pay, the "better" the benefits. Unfortunately, not everyone will obtain an upper-class status in life. This makes it incredibly difficult to obtain an insurance plan that can fulfill their needs. Not to mention; citizens with large families, the bill to foot for insurance is astronomically high. As a result, many lower income households and middle class households suffer from this current system. They have the choice of: pay rent and buy groceries or pay a ridiculously high premium in case something happens to them or a family member. The other option is to have no insurance at all, receive proper medical care, and then go into debt for many years paying off a hospital bill. There are many amazing solutions to this dystopian nightmare we have funneled ourselves into. Abolishing insurance companies and beginning a universal healthcare would be the most viable solution. A system where everyone chips in a little bit, usually in the form of a very small tax increase. Everyone at this point would receive the proper healthcare that they need. To add to this, universal basic income is another fascinating idea. The idea that all lower to middle income households will get a basic income to cover such necessities. Such as food, utility assistance, rent, and more. One day, we will all live in a utopian world where we learn that hoarding money and material possessions is not the definite goal for us as human beings. We will look back at this time in our history in embarrassment, as we look toward the stars for the advancement of the human race. There will be no corporations or systems set up to hurt the very people that built them. Working classes will be a thing of the past. The solutions for these complex issues will be set into motion and will work toward a brighter future. The question is, "when?". Well, that's up to you and I.

Model Rationales (Gemini vs. Claude)

Gemini Feedback:

The text demonstrates generally accurate English usage with a clear organizational structure (introduction, body paragraphs discussing the problem, potential solutions, and a conclusion). While there are minor errors such as "Unforunately" and occasional wordiness, these do not significantly impede comprehension. The vocabulary is appropriate, and the writer demonstrates a clear point of view on the healthcare system. The text exhibits stronger development and organization than the DevEd example, aligning more closely with the characteristics of college-level writing.

Claude Feedback:

This text exhibits characteristics of college-level writing in several key areas:

1. Mechanical conventions: The text demonstrates strong control over spelling, grammar, and punctuation. There are very few errors, and those present do not impede comprehension.

2. Sentence variety and style: The writing shows sophisticated sentence structure and varied sentence lengths. The vocabulary is advanced and appropriate for academic writing (e.g., "lucrative scams," "dystopian nightmare," "viable solution").

3. Development and support: The essay presents a clear argument about the problems with the current healthcare system and proposes solutions. It provides specific examples and elaborates on key points.

4. Organization and structure: The text follows a logical structure with an introduction, body paragraphs discussing different aspects of the issue, and a conclusion. Transitions between paragraphs are smooth and coherent.

5. Purpose and focus: The essay maintains a consistent focus on the healthcare system's problems and potential solutions. It addresses the topic with clarity and relevance to the intended audience.

6. Critical thinking: The writer demonstrates depth of thought by analyzing the healthcare system's impact on different social classes and proposing alternative systems. There are occasional minor issues (a few comma splices), but the overall quality of writing, complexity of ideas, and structure place this text at the College Level.

This page was left blank intentionally.