# 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/

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 reoccurring 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 $(\sigma)$ between the two arrays of ratios was also calculated to provide further insight into this relationship. A value of $(\sigma)$= 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 |
|---|---|---|---|---|---|---|---|
| Orthographic | <**CONTRACT**> | *417* | *6.42%* | *390* | *5.10%* | *807* | *5.71%* |
| | <**ORT**> | *1,504* | *23.16%* | *1,368* | *17.91%* | *2,872* | *20.32%* |
| | <**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% |
| Grammatical | <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% |
| | <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% |
| Lexical & Semantic | <CONNECT> | 100 | 1.54% | 215 | 2.81% | 315 | 2.23% |
| | <**MWE**> | *150* | *2.31%* | *281* | *3.68%* | *431* | *3.05%* |
| | <**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% |
| Discursive | <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% |
| | <**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.

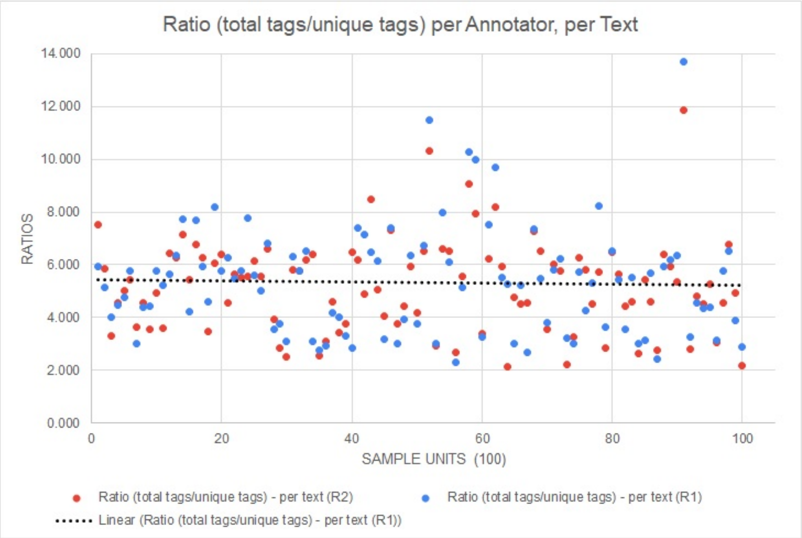

Figure 1: Ratio total tags/unique tags per annotators.

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 ($\sigma$) of the text units' K-alpha scores was also calculated, and a $\sigma$=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).

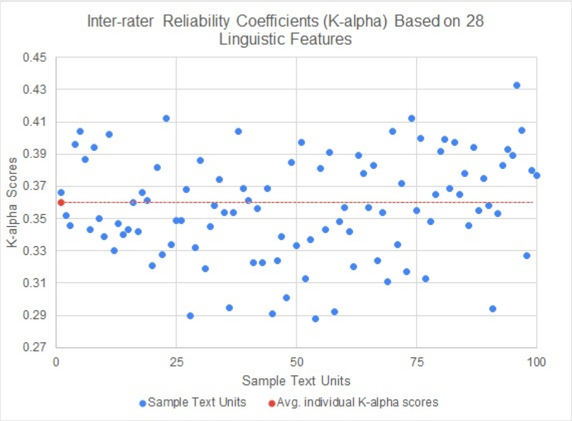

Figure 2: K-alpha scores' distribution around average value.

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 addressing 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.

## 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

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

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 postsecondary 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.