# Revisiting Classification Taxonomy for Grammatical Errors

Deqing Zou[1*], Jingheng Ye[1*], Yulu Liu[3], Yu Wu[1], Zishan Xu[1], Zihua Lan[1],
Yinghui Li[1], Hai-Tao Zheng[1,2†], Lan Zhou[4], Hong-Gee Kim [5]
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]University of Electronic Science and Technology of China, Chengdu, China
[4]Shenzhen Giiso Information Technology Co., Ltd., Shenzhen, China
[5]Seoul National University, Seoul, South Korea
{zdq23,yejh22}@mails.tsinghua.edu.cn

***Abstract*—Grammatical error classification plays a crucial role in language learning systems, but existing classification taxonomies are often adopted without rigorous validation, which can lead to inconsistencies and unreliable feedback. In this paper, we revisit previous classification taxonomies for grammatical errors by introducing a systematic and quantitative evaluation framework. Our approach examines four complementary aspects of taxonomy quality, namely exclusivity, coverage, consistency, and usability. To support this evaluation, we construct a high-quality grammatical error classification dataset annotated with multiple classification taxonomies and evaluate them under our proposed evaluation framework. Experimental results reveal clear drawbacks and limitations of existing taxonomies. Overall, our contributions aim to improve the precision and effectiveness of error analysis, providing more understandable and actionable feedback for language learners.**

***Index Terms*—Classification Taxonomy, Evaluation Framework, Large Language Model, Grammatical Error Classification**

## I. Introduction

Errors are an inevitable aspect of language acquisition, serving as critical indicators of learners' linguistic development and providing valuable insights for educators and intelligent language learning systems [1]–[4]. In Error Analysis (EA), the systematic identification, categorization, and interpretation of learner errors play a pivotal role in improving personalized instruction, generating automated feedback, and enabling effective language assessment [5]–[8]. Central to this process is the use of grammatical error classification taxonomies [9]–[11], which organize learner errors according to linguistic or cognitive principles. These taxonomies have been widely adopted in applications such as grammatical error correction (GEC) [12]–[14] and automated essay scoring (AES), significantly enhancing error detection [15], [16], correction [17], and feedback generation [18], [19].

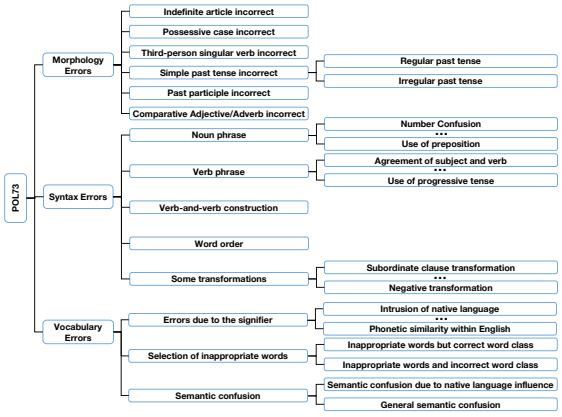
A well-designed grammatical error classification taxonomy allows language learners to better understand the nature and causes of their errors, facilitating targeted improvements in their linguistic competence [20]. However, existing taxonomies are often developed based on empirical assumptions or ad-hoc practices, without rigorous validation [21], [22]. This lack of systematic evaluation has led to *issues* such as overlapping categories, insufficient coverage of error types, and limited applicability in real-world educational contexts.

To address these challenges, this paper revisits grammatical error classification taxonomies by systematically assessing their quality and utility. Specifically, we introduce a multi-metric evaluation framework that examines four key dimensions: ❶ *Exclusivity* ensures that error categories are mutually exclusive, with clearly defined boundaries to minimize overlap and ambiguity; ❷ *Coverage* measures the extent to which a taxonomy captures both common and rare error types, ensuring comprehensive representation; ❸ *Consistency* assesses whether instances within the same category exhibit coherent and stable semantic characteristics, reflecting internal category coherence and clear distinction from other categories rather than fragmented groupings; ❹ *Usability* evaluates how reliably the taxonomy can be applied in both human annotation and model-based classification settings.

To validate our evaluation framework, we construct a high-quality grammatical error dataset annotated with multiple classification taxonomies. The dataset is created through a collaborative annotation process combining large language models (LLMs) and human annotators, ensuring both scalability and annotation reliability. Using this dataset, we systematically evaluate four widely used error classification taxonomies—**POL73** [23], **TUC74** [24], **BRY17** [22], and **FEI23** [11]—which represent diverse design philosophies and category structures (see Fig. 1). Through performance comparisons, annotator agreement analyses, and an ablation study on error type merging, we demonstrate the necessity of systematic evaluation for grammatical error classification taxonomies. In summary, our contributions are as follows:

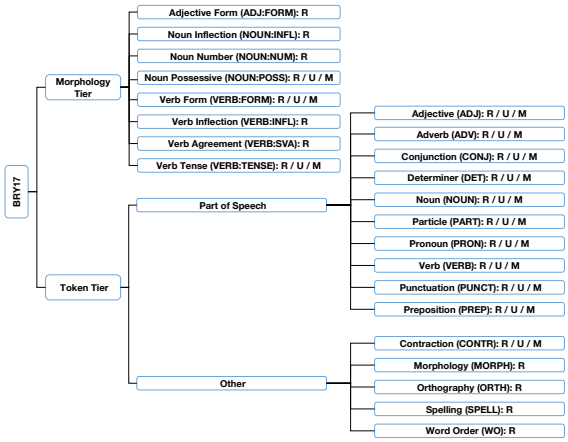
- We propose a novel multi-metric evaluation framework for grammatical error classification taxonomies, spanning exclusivity, coverage, consistency, and usability.
- We construct a high-quality grammatical error dataset annotated with multiple taxonomies, leveraging a collaborative

*Equal contribution.
†Corresponding author: Hai-Tao Zheng. (Email: zheng.haitao@sz.tsinghua.edu.cn)

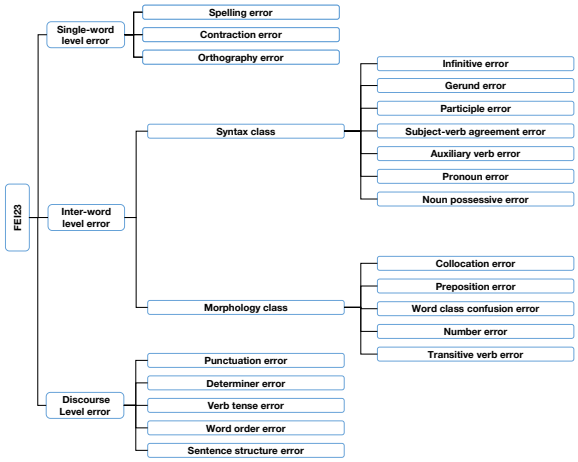



(a) POL73 Error Classification Taxonomy


(b) TUC74 Error Classification Taxonomy




(c) BRY17 Error Classification Taxonomy




(d) FEI23 Error Classification Taxonomy

Fig. 1: Overview of four representative error classification taxonomies. Each subfigure visualizes the category organization.

annotation process involving LLMs and human experts.

- We conduct a comprehensive empirical comparison of four widely used taxonomies, revealing systematic trade-offs in their design and practical applicability.

## II. Related Work

Grammatical error classification has long been studied in second language acquisition, learner corpus research, and grammatical error correction. Prior work has proposed diverse taxonomies motivated by linguistic description, surface error patterns, cognitive perspectives, and practical annotation needs. Representative examples include linguistically grounded taxonomies such as POL73 [23] and TUC74 [24], surface-operation-based schemes such as ERRANT [22], and cognitively motivated frameworks like FEI23 [11], as well as corpus-driven annotation frameworks including the Cambridge Learner Corpus [25] and CLEME [26]. In this work, we focus on four taxonomies widely used in empirical studies, annotation practice, and benchmark evaluation.

**POL73** [23] classifies errors by affected linguistic components (e.g., morphology, syntax, and vocabulary), representing an early descriptive, component-based taxonomy.

**TUC74** [24] defines a hierarchical taxonomy grounded in English sentence structure, modeling errors as violations of syntactic dependencies and clause-level relations.

**BRY17** [22], implemented in the ERRANT framework, adopts a surface-operation-based taxonomy combining edit operations and part-of-speech categories.

**FEI23** [11] proposes a cognitively motivated taxonomy that organizes grammatical errors across different levels of linguistic processing, enabling interpretable error analysis.

Despite their widespread adoption, these taxonomies differ substantially in category granularity, conceptual boundaries, and operational definitions, making analytical conclusions highly taxonomy-dependent. However, most existing studies adopt them from prior work with little explicit scrutiny of their structural assumptions, boundary clarity, or practical reliability, motivating the need for a systematic evaluation framework for controlled and comparable taxonomy analysis.

## III. Methodology

Given a dataset $D$ of English learner texts, our goal is to quantitatively evaluate $n$ error classification taxonomies $\{F_1, F_2, \ldots, F_n\}$, where each taxonomy $F_i$ consists of $m$

predefined error types $\{ET_1, \ldots, ET_m\}$. We evaluate taxonomies along four complementary dimensions: **Exclusivity**, **Coverage**, **Consistency**, and **Usability**. Exclusivity and Usability are evaluated based on error classification outputs produced by large language models. Coverage is computed from a manually annotated dataset via statistical analysis. Consistency is assessed by measuring the semantic coherence and separability of error types in an embedding space.

### A. Exclusivity

The error types in a taxonomy should be mutually exclusive, meaning that each error instance belongs to a single category. Overlapping error types introduce instability and inconsistency in error analysis, reducing classification reliability. In practice, if a model assigns high confidence to multiple categories for the same instance, it indicates that the taxonomy lacks clear boundaries. To quantify exclusivity, we analyze LLM-generated confidence scores. Since raw estimates are often noisy or overconfident, we adopt a robust estimation scheme that combines three complementary strategies: (1) *Top-$k$ prompting*, which elicits multiple plausible error types with confidence scores; (2) *self-random sampling*, which repeats the same prompt under fixed decoding settings (e.g., temperature and top-$p$) to estimate uncertainty from stochastic sampling within the truncated probability distribution; and (3) *average-confidence aggregation*, which averages confidence scores across multiple samples to improve stability [27].

Specifically, a sample is considered classified under an error type if its confidence score exceeds the predefined threshold $\tau$. Following prior work on confidence elicitation, we set $\tau = 0.7$, which provides a conservative balance between filtering low-confidence predictions and retaining sufficient coverage across models. We define the $\mathrm{Overlap}$ to quantify the degree of category overlap for a sample $x$ as follows:

$$\mathrm{Overlap}(x) = |\{\hat{Y}_i^{(x)} \mid C_i^{(x)} > \tau\}|, \tag{1}$$

where $\hat{Y}_i^{(x)}$ denotes the $i$-th predicted error type for a given sample $x$, and $C_i^{(x)}$ represents its confidence score. A higher $\mathrm{Overlap}(x)$ indicates that multiple error types are assigned to the same sample, suggesting a violation of exclusivity.

The *Exclusivity Score* is computed as the average instance-level exclusivity over the dataset $D$:

$$\mathrm{Exclusivity}(F) = \frac{1}{|D|} \sum_{x \in D} \begin{cases} 1 - \frac{\mathrm{Overlap}(x)-1}{k-1}, & \mathrm{Overlap}(x) > 0, \\ 0, & \mathrm{Overlap}(x) = 0, \end{cases} \tag{2}$$

where $k$ denotes the number of candidate error types elicited by the Top-$k$ prompting strategy, which is fixed to $k = 3$ in all experiments. $\mathrm{Exclusivity}$ indicates whether the classification taxonomy maintains clear distinctions between error types. A lower score indicates substantial category overlap and ambiguous boundaries, whereas a higher score reflects clearer and more reliable taxonomic distinctions

### B. Coverage

Coverage measures the proportion of error instances in $D$ that can be assigned to predefined categories. Let $|U|$ denote the number of such errors (excluding "Other") and $|D|$ the total number of error instances. The *Coverage Score* is defined as:

$$\mathrm{Coverage}(F) = \frac{|U|}{|D|}. \tag{3}$$

A higher $\mathrm{Coverage}$ indicates that the taxonomy captures a larger proportion of observed errors, reflecting greater completeness of error type representation.

### C. Consistency

A high-quality error classification taxonomy should exhibit both strong internal coherence within each error type and clear separation between different error types, ensuring clear category boundaries. To quantify these complementary properties, we define *Intra-Type Consistency* (ITC) and *Inter-Type Separability* (ITS) based on embedding-space representations of error instances. Specifically, we represent each error instance in a continuous embedding space by mapping erroneous and corrected sentences into a shared semantic space using the pretrained *Qwen3-Embedding-8B* model [28]. Given an erroneous sentence $x_{\text{error}}$ and its corrected counterpart $x_{\text{correct}}$, we define the *error embedding* of a sample $x$ as:

$$v(x) = \mathrm{Embedding}(x_{\text{correct}}) - \mathrm{Embedding}(x_{\text{error}}), \tag{4}$$

which explicitly encodes the transformation from the erroneous form to the corrected form. All error embeddings are L2-normalized prior to similarity computation.

Let $D_j$ denote the set of error instances associated with the $j$-th error type $ET_j$ in taxonomy $F$. The centroid of error type $ET_j$ is defined as:

$$\mu_j = \frac{1}{|D_j|} \sum_{x \in D_j} v(x), \tag{5}$$

followed by L2-normalization, such that $\mu_j$ represents the mean direction of error embeddings for $ET_j$.

*a) Intra-Type Consistency:* Intra-type consistency measures the compactness of error instances within the same error type. For each $ET_j$, we compute:

$$\mathrm{ITC}_j = \begin{cases} \frac{1}{|D_j|} \sum_{x \in D_j} \cos\big(v(x), \mu_j\big), & |D_j| > 1, \\ 0, & |D_j| = 1, \end{cases} \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

The overall *Intra-Type Consistency* of taxonomy $F$ is defined as a weighted average over all error types:

$$\mathrm{ITC}(F) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \mathrm{ITC}_j \,. \tag{7}$$

*b) Inter-Type Separability:* Inter-type separability evaluates the separability between different error types based on the pairwise distances between their centroid embeddings. Given the centroids $\{\mu_j\}_{j=1}^{m}$ of all error types in taxonomy $F$, the distance between two error types $ET_i$ and $ET_j$ is defined as:

$$d(i, j) = 1 - \cos\big(\mu_i, \mu_j\big), \tag{8}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

The global *Inter-Type Separability* of taxonomy $F$ is then computed as the average pairwise distance:

$$\mathrm{ITS}(F) = \frac{2}{m(m-1)} \sum_{1 \le i < j \le m} d(i,j). \tag{9}$$

*c) Interpretation:* By normalizing both error embeddings and centroids, the proposed metrics capture taxonomy consistency in terms of structural coherence based solely on directional similarity. Higher $\mathrm{ITC}(F)$ indicates more compact intra-type clusters, while higher $\mathrm{ITS}(F)$ reflects clearer separation between error types. Together, these metrics quantitatively characterize taxonomy structure in terms of internal coherence and inter-type separation.

### D. Usability

We argue that a usable error taxonomy should be effective for both automated models and human annotators. Accordingly, we assess usability from two complementary perspectives: *model-level usability* and *human-level usability*.

*a) Model-level Usability:* Model-level usability reflects whether a taxonomy can be reliably learned and applied by LLMs, supporting robust automatic error analysis. We quantify this aspect using *Macro F1* and *Micro F1* scores:

$$\begin{aligned} \mathrm{Macro\_F1} &= \frac{1}{m} \sum_{i=1}^{m} \frac{2 \cdot P(ET_i) \cdot R(ET_i)}{P(ET_i) + R(ET_i)}, \\ \mathrm{Micro\_F1} &= \frac{2 \cdot P(D) \cdot R(D)}{P(D) + R(D)}, \end{aligned} \tag{10}$$

where $P(\cdot)$ and $R(\cdot)$ denote precision and recall, respectively. A higher Macro F1 indicates that the taxonomy enables balanced model performance across both frequent and infrequent error types, while a higher Micro F1 reflects robustness under large-scale error detection. Together, these metrics characterize the effectiveness of a taxonomy in supporting stable and scalable model-based error classification.

*b) Human-level Usability:* Human-level usability captures how intuitively and consistently a taxonomy can be applied by human annotators. A usable taxonomy should minimize ambiguity in category definitions and yield high inter-annotator agreement. We therefore assess this aspect using inter-annotator agreement, as detailed in Section V-C.

## IV. Experimental Settings

### A. Dataset and Annotation

*1) Dataset Overview:* We use a dataset derived from the Cambridge English Write & Improve (W&I) and LOCNESS corpora [29], re-annotated using ERRANT [22]. After preprocessing, the dataset contains 487 instances, each corresponding to a single grammatical error. Each instance is annotated under four representative taxonomies: POL73, TUC74, BRY17, and FEI23. All data are anonymized and contain no sensitive personal or identifiable information of any kind.

*2) Data Preprocessing:* Sentences containing multiple grammatical errors may introduce ambiguity in classification. To address this, each sentence with $n > 1$ errors is decomposed into $n$ instances, each retaining one error while correcting the others. This process is implemented using ERRANT [22] with CLEME [26], [30]. As a result, each instance contains a single well-defined error, enabling more reliable and interpretable classification.

*3) Annotation Approach:* We adopt a collaborative annotation framework combining LLMs with human expertise [31]. In the first stage, an LLM generates preliminary annotations, which are refined by the authors through prompt optimization to align with taxonomy definitions. In the second stage, two professional linguists independently re-annotate the data, and disagreements are resolved through discussion with the authors to reach consensus. This multi-stage process ensures reliable and consistent annotations across taxonomies.

### B. Model Details

We evaluate different grammatical error taxonomies using a diverse set of state-of-the-art LLMs, including **DeepSeek-V3.2** [32], **Qwen3-235B-A22B-Instruct** [33], **Doubao-1.5-Pro** [34], **GPT-5-mini** [35], **GPT-4o** [36], **Claude-Sonnet-4.5** [37], and **Grok-4-Fast** [38]. These models exhibit strong capabilities in text understanding, reasoning, and fine-grained classification. All models are evaluated under a unified task formulation and output format defined by our framework.

## V. Experimental Results

### A. Comparative Analysis of Taxonomies

We systematically assess the rationality of different error classification taxonomies along four complementary dimensions—**Exclusivity**, **Coverage**, **Consistency**, and **Usability**—as summarized in Table I. The results reveal that systematic design differences lead to distinct empirical behaviors, closely tied to how error types are defined and operationalized.

**Exclusivity: Ambiguous boundaries undermine mutual exclusivity.** Taxonomies with overlapping linguistic criteria exhibit lower exclusivity, as errors may plausibly fit multiple categories. POL73 and TUC74 contain conceptually adjacent or partially redundant categories, leading to frequent ambiguity. In contrast, BRY17 defines error types via edit operations and part-of-speech constraints, enforcing a single dominant interpretation and achieving higher exclusivity.

**Coverage: Surface-oriented categories improve empirical coverage.** Coverage depends on whether frequent surface-level error phenomena are explicitly modeled. POL73 and TUC74 lack categories for spelling, orthography, and punctuation, resulting in lower coverage. By contrast, BRY17 and FEI23 explicitly include such errors through surface-oriented or word-level categories, achieving substantially higher coverage. This indicates that empirical completeness depends more on modeling frequent errors than on theoretical granularity.

**Consistency: Structural granularity shapes coherence.** Consistency reflects a trade-off between ITC and ITS. Fine-grained taxonomies such as BRY17 achieve strong intra-

TABLE I: Performance comparison of classification taxonomies. Higher values indicate better performance in all metrics.

| Model | Exclusivity ↑ | | | | Coverage ↑ | | | | Consistency (ITC / ITS) ↑ | | | | Usability (Macro / Micro F1) ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | POL73 | TUC74 | BRY17 | FEI23 | POL73 | TUC74 | BRY17 | FEI23 | POL73 | TUC74 | BRY17 | FEI23 | POL73 | TUC74 | BRY17 | FEI23 |
| DeepSeek-V3.2 | 0.699 | 0.393 | **0.737** | 0.655 | | | | | | | | | 0.333 / 0.487 | 0.068 / 0.082 | 0.573 / **0.780** | **0.595** / 0.741 |
| Qwen3-235B-A22B-Instruct | 0.794 | 0.533 | **0.861** | 0.733 | | | | | | | | | 0.318 / 0.448 | 0.103 / 0.090 | **0.530** / **0.758** | 0.518 / 0.731 |
| Doubao-1.5-Pro | 0.569 | 0.208 | **0.709** | 0.668 | | | | | | | | | 0.093 / 0.351 | 0.043 / 0.092 | 0.228 / 0.530 | **0.302** / **0.690** |
| GPT-5-mini | 0.752 | 0.250 | **0.928** | 0.853 | 0.698 | 0.160 | **0.980** | 0.924 | 0.215 / 0.894 | 0.158 / **0.950** | **0.288** / 0.931 | 0.221 / 0.824 | 0.562 / 0.534 | 0.315 / 0.119 | 0.626 / **0.821** | **0.645** / 0.772 |
| GPT-4o | 0.535 | 0.463 | **0.632** | 0.471 | | | | | | | | | 0.291 / 0.472 | 0.048 / 0.092 | 0.565 / **0.739** | **0.598** / 0.704 |
| Claude-Sonnet-4.5 | 0.697 | 0.219 | **0.796** | 0.771 | | | | | | | | | 0.507 / 0.546 | 0.224 / 0.099 | **0.750** / **0.850** | 0.678 / 0.799 |
| Grok-4-Fast | 0.630 | 0.184 | **0.760** | 0.640 | | | | | | | | | 0.486 / 0.542 | 0.278 / 0.107 | 0.643 / **0.819** | **0.668** / 0.799 |
| **Average** | 0.668 | 0.321 | **0.775** | 0.685 | 0.698 | 0.160 | **0.980** | 0.924 | 0.215 / 0.894 | 0.158 / **0.950** | **0.288** / 0.931 | 0.221 / 0.824 | 0.370 / 0.483 | 0.154 / 0.097 | 0.559 / **0.757** | **0.572** / 0.748 |

TABLE II: Ablation Study on the Impact of Merging Error Categories (GPT-5-mini). Taxonomies labeled with Fusion indicate error classification taxonomies after category merging.

| Taxonomies | Exclusivity ↑ | Coverage ↑ | Consistency (ITC/ITS) ↑ | Usability (Macro/Micro F1) ↑ |
|---|---|---|---|---|
| POL73 | **0.752** | 0.698 | **0.215/0.894** | **0.562/0.534** |
| POL73 (Fusion) | 0.743 | **0.723** | 0.212/0.889 | 0.443/0.501 |
| TUC74 | 0.250 | 0.160 | **0.158/0.950** | **0.315/0.119** |
| TUC74 (Fusion) | **0.414** | **0.405** | **0.158**/0.942 | 0.256/0.103 |
| BRY17 | **0.928** | 0.980 | **0.288/0.931** | **0.626/0.821** |
| BRY17 (Fusion) | 0.900 | **0.982** | 0.284/0.927 | 0.422/0.667 |
| FEI23 | 0.853 | 0.924 | **0.221**/0.824 | **0.645/0.772** |
| FEI23 (Fusion) | **0.859** | **0.986** | 0.213/**0.830** | 0.584/0.711 |

TABLE III: Cohen's Kappa Scores for Inter-Annotator Agreement across taxonomies.

| Annotator Pair | POL73 | TUC74 | BRY17 | FEI23 |
|---|---|---|---|---|
| Annotator 1 & 2 | 0.691 | 0.692 | **0.808** | 0.714 |
| Annotator 2 & 3 | 0.813 | 0.661 | **0.851** | 0.824 |
| Annotator 1 & 3 | 0.636 | 0.547 | **0.734** | 0.651 |
| Average | 0.713 | 0.633 | **0.798** | 0.730 |

type consistency through similar surface edit patterns, but show moderate inter-type separability due to closely related categories. Conversely, TUC74 achieves stronger inter-type separability via structural organization, at the cost of reduced intra-type consistency. This suggests that consistency depends on balancing surface specificity and structural abstraction.

**Usability: Operational clarity outweighs theoretical expressiveness.** Taxonomies requiring abstract linguistic judgments are harder to apply consistently in practice. POL73 and TUC74 rely on fine-grained linguistic interpretation, often reducing annotation reliability. In contrast, BRY17 and FEI23 adopt operationally transparent definitions aligned with surface edits or processing levels, enabling more consistent annotation and higher Macro and Micro F1 scores overall. This highlights operational clarity as a key determinant of usability in practice.

Overall, these findings show that taxonomies encode fundamentally different assumptions about error types. Operation-based and cognitively motivated taxonomies demonstrate stronger empirical robustness, while linguistically detailed taxonomies provide richer descriptions at the cost of lower exclusivity, consistency, and usability. These results highlight the need for principled, quantitative evaluation of taxonomy quality, as proposed in this work.

### *B. Ablation Study: Impact of Error Category Fusion*

To further validate the proposed evaluation metrics, we conduct an ablation study by selectively merging error categories within each taxonomy and examining the impact of reduced granularity on **Exclusivity**, **Coverage**, **Consistency**, and **Usability**. Results are summarized in Table II.

For POL73, merging verb-phrase subcategories into a coarse *Verb Phrase* category slightly improves coverage but degrades exclusivity, consistency, and usability, indicating increased ambiguity from over-aggregation. For TUC74, consolidating omission-related categories into *Missing Parts* improves coverage and exclusivity, while consistency and usability show limited gains, suggesting residual internal heterogeneity. For BRY17, merging function-word categories reduces exclusivity, consistency, and usability with minimal change in coverage, highlighting the sensitivity of its fine-grained, operation-based design. For FEI23, merging spelling, contraction, and orthography errors into a *Single-word Level* category increases coverage but reduces consistency and usability, indicating the importance of fine-grained distinctions.

Overall, category fusion tends to improve coverage but often harms exclusivity, consistency, and usability by increasing intra-category heterogeneity. Taxonomies that are already well balanced in granularity and operational clarity, such as BRY17, are particularly sensitive to fusion. These results show that the proposed metrics respond predictably to controlled structural changes, supporting their validity.

### *C. Annotator Agreement Analysis*

We evaluate inter-annotator agreement (IAA) as part of **Usability** using Cohen's Kappa [39]:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{11}$$

where $p_o$ is the observed agreement and $p_e$ is the expected agreement by chance. We randomly sample 100 instances and ask three graduate students in English linguistics to independently assign one error type per taxonomy. Pairwise Kappa scores are computed, and results are reported in Table III. Results show a consistent ranking across taxonomies: BRY17 achieves the highest agreement, followed by FEI23 and POL73, while TUC74 performs worst. This trend aligns with usability and overall performance in Table I, suggesting that taxonomies easier for humans to apply also perform better, supporting the validity of the usability metric.

## VI. Conclusion

In this work, we propose a systematic framework for evaluating grammatical error classification taxonomies along four complementary dimensions: **Exclusivity**, **Coverage**, **Consistency**, and **Usability**. These metrics operationalize key principles of effective taxonomies, including clear category boundaries, sufficient coverage, semantic coherence, and practical applicability. Experiments on a multi-taxonomy annotated dataset show that widely used taxonomies differ substantially across these dimensions, revealing trade-offs between structural granularity and operational clarity. We further show that usability is a coherent and informative property, consistently aligned with both model performance and human annotation agreement. These findings remain stable across multiple LLMs, indicating that the proposed framework provides robust and model-agnostic insights. While our approach leverages LLMs and embedding-based representations, interpretability is addressed at the taxonomy level by focusing on structural properties rather than model-specific behavior. Limitations include the relatively small dataset size, potential data contamination from public corpora, and the focus on a limited set of English taxonomies. Overall, this work highlights the importance of multi-dimensional taxonomy evaluation and provides a principled foundation for future taxonomy design, with potential extensions to broader classification systems.

## Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154); the Natural Science Foundation of Guangdong Province (Grant No.2024TQ08X729); Basic Research Fund of Shenzhen City (Grant No.JCYJ20240813112009013 and GJHZ20240218113603006); The Major Key Project of PCL for Experiments and Applications (Grant No.PCL2024A08).

## References

[1] H. C. Dulay, "Goofing: An indicator of children's second language learning strategies," *Language Learning*, 1972.
[2] M. Dodigovic, "Artificial intelligence and second language learning: An efficient approach to error remediation," *Language Awareness*, 2007.
[3] T. Heift and M. Schulze, *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge, 2007.
[4] Y. Li, Z. Xu *et al.*, "Towards real-world writing assistance: A Chinese character checking benchmark with faked and misspelled characters," in *ACL*, 2024, pp. 8656–8668.
[5] S. P. Corder, "The significance of learners' errors," *International Review of Applied Linguistics*, 1967.
[6] C. James, *Errors in Language Learning and Use: Exploring Error Analysis*, 1st ed. Routledge, 1998.
[7] E. Bialystok, H. Dulay, M. Burt, and S. Krashen, "Language two," *The Modern Language Journal*, vol. 67, p. 273, 1982.
[8] Y. Li, H. Huang *et al.*, "On the (in)effectiveness of large language models for chinese text correction," *arXiv preprint:2307.09007*, 2023.
[9] S. Ma, Y. Li *et al.*, "Linguistic rules-based corpus generation for native Chinese grammatical error correction," in *EMNLP*, 2022, pp. 576–589.
[10] J. Ye, S. Qin *et al.*, "Excgec: A benchmark of edit-wise explainable chinese grammatical error correction," 2024.
[11] Y. Fei, L. Cui, S. Yang *et al.*, "Enhancing grammatical error correction systems with explanations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 7489–7501.
[12] J. Ye, Y. Li *et al.*, "Mixedit: Revisiting data augmentation and beyond for grammatical error correction," in *EMNLP*, 2023, pp. 10 161–10 175.
[13] J. Ye, Y. Li, and H.-T. Zheng, "System report for ccl23-eval task 7: Thu kelab (sz)," in *Proceedings of CCL*, 2023, pp. 262–270.
[14] Y. Li *et al.*, "Correct like humans: Progressive learning framework for chinese text error correction," *Expert Systems with Applications*, 2025.
[15] H. Huang, J. Ye *et al.*, "A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check," in *EMNLP*, 2023.
[16] D. Zhang, Y. Li, Q. Zhou, and S. Ma, "Contextual similarity is more valuable than character similarity: An empirical study for chinese spell checking," in *Proceedings of ICASSP*, 2023.
[17] J. Ye, Y. Li *et al.*, "Focus is what you need for chinese grammatical error correction," *arXiv preprint:2210.12692*, 2022.
[18] K.-H. Liang, S. Davidson, X. Yuan *et al.*, "ChatBack: Investigating methods of providing grammatical error feedback in a gui-based language learning chatbot," in *Proceedings of the BEA*, 2023.
[19] H. Yannakoudakis and M. Rei, "Neural sequence-labelling models for grammatical error correction," in *EMNLP*, 2017.
[20] J. Ye, S. Wang, D. Zou *et al.*, "Position: Llms can be good tutors in foreign language education," *arXiv preprint:2502.05467*, 2025.
[21] J. He, B. Peng, Y. Liao, Q. Liu *et al.*, "Tgea: An error-annotated dataset and benchmark tasks for text generation from pretrained language models," in *ACL*, 2021.
[22] C. Bryant, M. Felice *et al.*, "Automatic annotation and evaluation of error types for grammatical error correction," in *ACL 2017*, 2017.
[23] R. L. Politzer and A. G. Ramirez, "An error analysis of the spoken english of mexican-american pupils in a bilingual school and a monolingual school," *Language Learning*, vol. 23, no. 1, pp. 39–61, 1973.
[24] G. R. Tucker, M. K. Burt, and C. Kiparsky, "The gooficon: A repair manual for english," *TESOL Quarterly*, vol. 8, no. 2, p. 191, 1974.
[25] D. Nicholls, "The cambridge learner corpus: Error coding and analysis for lexicography and elt," in *Proceedings of Corpus Linguistics*, 2003.
[26] J. Ye, Y. Li *et al.*, "Cleme: Debiasing multi-reference evaluation for grammatical error correction," in *EMNLP*, 2023.
[27] M. Xiong, Z. Hu *et al.*, "Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms," *arXiv preprint arXiv:2306.13063*, 2024, accepted to ICLR 2024.
[28] Y. Zhang and M. Li, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," *arXiv preprint arXiv:2506.05176*, 2025.
[29] C. Bryant and M. Felice, "The bea-2019 shared task on grammatical error correction," in *Proceedings of BEA*, 2019.
[30] J. Ye, Z. Xu *et al.*, "CLEME2.0: Towards interpretable evaluation by disentangling edits for grammatical error correction," in *ACL*, 2025.
[31] P. Törnberg, "Best practices for text annotation with large language models," *Sociologica*, vol. 18, no. 2, pp. 67–85, 2024.
[32] DeepSeek-AI, "Deepseek-v3.2: Pushing the frontier of open large language models," *arXiv preprint:2512.02556*, 2025.
[33] Q. Team, "Qwen3 technical report," *arXiv preprint:2505.09388*, 2025.
[34] ByteDance Seed, "Doubao-1.5-pro," https://seed.bytedance.com/en/special/doubao_1_5_pro, 2025, accessed: Jan. 22, 2025.
[35] OpenAI, "Introducing GPT-5," https://openai.com/index/introducing-gpt-5/, 2024, accessed: Aug. 7, 2025.
[36] OpenAI, "Gpt-4o system card," *arXiv preprint:2410.21276*, 2024.
[37] Anthropic, "Introducing claude sonnet 4.5," https://www.anthropic.com/news/claude-sonnet-4-5, 2025, published: Sept. 30, 2025.
[38] x.ai, "Grok 4 fast: Pushing the frontier of cost-efficient intelligence," https://x.ai/news/grok-4-fast, 2025, published: Sept. 19, 2025.
[39] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.