Revisiting Classification Taxonomy for Grammatical Errors

Anonymous ACL submission

Abstract

Grammatical error classification plays a crucial role in language learning systems, but existing classification taxonomies often lack rigorous validation, leading to inconsistencies and unreliable feedback. In this paper, we revisit previous classification taxonomies for grammatical errors by introducing a systematic and qualitative evaluation framework. Our approach examines four aspects of a taxonomy, i.e., exclusivity, coverage, balance, and usability. Then, we construct a high-quality grammatical error classification dataset annotated with multiple classification taxonomies and evaluate them grounding on our proposed evaluation framework. Our experiments reveal the drawbacks of existing taxonomies. Our contributions aim to improve the precision and effectiveness of error analysis, providing more understandable and actionable feedback for language learners.[1]

1 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 (Dulay and Burt, 1972; Dodigovic, 2007; Heift and Schulze, 2007). 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 (Corder, 1967; James, 1998; Bialystok et al., 1982). Central to this process is the use of grammatical error classification taxonomies (Ye et al., 2024; Fei et al., 2023), which organize learner errors based on linguistic or cognitive principles. These taxonomies have been widely adopted in applications such as grammatical error correction (GEC) and automated essay scoring (AES), significantly enhancing error detection, correction, and feedback generation (Liang et al., 2023; Yannakoudakis et al., 2017).

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. However, existing taxonomies are often developed based on empirical assumptions or ad-hoc practices, without rigorous validation (He et al., 2021; Bryant et al., 2017). 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 of a taxonomy: ❶ *Exclusivity* ensures that error categories are mutually exclusive, with clearly defined boundaries to minimize overlap and ambiguity; ❷ *Coverage* evaluates the extent to which the taxonomy captures both common and rare error types, ensuring a comprehensive representation of learner errors. ❸ *Balance* measures the taxonomy's ability to balance attention between frequent and infrequent error types, avoiding overemphasis on a narrow subset of errors; ❹ *Usability* assesses the clarity and practical applicability of the taxonomy.

To validate our evaluation framework, we construct a high-quality grammatical error dataset annotated with multiple classification taxonomies. This dataset is created through a collaborative annotation approach that leverages large language models (LLMs) and human annotators, ensuring scalability and annotation reliability. Using this dataset, we systematically evaluate four widely-used error classification taxonomies: **POL73** (Politzer and Ramirez, 1973), **TUC74** (Tucker et al., 1974), **BRY17** (Bryant et al., 2017), and **FEI23** (Fei et al.,

[1] All the codes and data will be released after the review.

2023). Through performance comparisons and annotator agreement experiments, we evaluate the rationality of these taxonomies. An ablation study on error type merging further reveals that classification taxonomies should not rely solely on empirical intuition but require systematic validation. These findings underscore the necessity of a rigorous evaluation for grammatical error taxonomy. In summary, our contributions are as follows:

- We propose a novel multi-metric evaluation framework for grammatical error classification taxonomies, incorporating dimensions of exclusivity, coverage, balance, and usability.
- We construct a high-quality grammatical error dataset annotated with multiple taxonomies, leveraging a collaborative annotation process involving LLMs and human experts.
- We conduct a comprehensive evaluation of four widely used taxonomies, providing insights into their strengths, limitations, and practical implications for error analysis in language learning.

2 Related Work

Existing studies propose various error classification methods based on different perspectives, such as linguistic structure, cognitive processes, and communicative impact. Bialystok et al. (1982) categorize language errors into four primary taxonomies: (1) *Linguistic Category Taxonomy* (Politzer and Ramirez, 1973; Tucker et al., 1974; Ângela Costa et al., 2015) classifies errors based on language components or linguistic constituents; (2) *Surface Strategy Taxonomy* (Özkayran and Yılmaz, 2020; Rixha et al., 2021; Suhono, 2017) focuses on structural modifications made by learners; (3) *Comparative Taxonomy* (Kafipour and Laleh, 2012; Dulay and Burt, 1974; White and Ontario Institute for Studies in Education, 1977) classifies errors by comparing L2 learner errors to L1 acquisition errors or native language structures; (4) *Communicative Effect Taxonomy* (Burt, 1975; Waruwu and Harefa, 2024) classify errors based on their communicative impact. In addition to the above systematic classification taxonomies, other studies (CHAN, 2010; Dahlmeier et al., 2013; Nicholls, 2003; Fei et al., 2023) have introduced other taxonomies, such as the simplification strategy taxonomy (Bertkua, 1974) and the rule-based annotation toolkit (Bryant et al., 2017).

3 Methodology

Given a dataset D of English learner texts, our goal is to systematically and quantitatively evaluate the rationality of n error classification taxonomies $F = \{F_1, F_2, ..., F_n\}$. Each taxonomy F_i consists of m predefined error types, denoted as $F_i = \{ET_1, ET_2, ..., ET_m\}$, where ET_j represents the j-th error type in the taxonomy F_i. All notations are detailed in the Appendix A.

We evaluate the rationality of each taxonomy along four dimensions: **Exclusivity**, **Coverage**, **Balance**, and **Usability**. Exclusivity and Usability are based on the generated results by the LLM, while Balance and Coverage are computed obtained directly from the manually annotated dataset through statistical analysis.

3.1 Exclusivity

The error types in the taxonomy should be mutually exclusive, meaning that each error instance belongs to a single distinct category. Overlapping error types introduce instability and inconsistencies in error analysis, reducing the reliability of classification results. If a model frequently assigns high confidence to multiple categories for the same error, it suggests that the taxonomy lacks clear boundaries between certain error types. To quantify this issue, we assess exclusivity by analyzing the confidence scores of an LLM's predictions. Confidence estimation plays a crucial role in this evaluation, we incorporate three established methods (Xiong et al., 2024) for improved reliability (details are in Appendix C). Specifically, a sample is considered classified under an error type if its confidence score exceeds the predefined threshold τ. We define the $\mathrm{Overlap}$ to quantify the degree of category overlap for a sample x as follows:

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

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:

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

where k represents the selection parameter in the Top-K Prompting Strategy, detailed in Appendix C. Exclusivity indicates whether the classification taxonomy maintains clear distinctions between error types. A lower score suggests significant overlap between categories, indicating poorly defined boundaries, whereas a higher score implies a more distinct and reliable classification system.

3.2 Coverage

Coverage measures the extent to which a taxonomy accounts for errors in the dataset D. Let $|U|$ be the number of errors labeled with at least one defined category (as opposed to "Other"), and $|D|$ be the total number of error instances. We define *Coverage score* as follows:

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

A higher Coverage indicates that the taxonomy captures a greater proportion of actual errors, demonstrating superior completeness in covering the range of error types.

3.3 Balance

Error classification taxonomy should maintain a balanced distribution of error types, avoiding excessive concentration on a few dominant categories while ensuring sufficient representation of less frequent errors. An imbalanced taxonomy may exhibit a long-tail effect (Zhang et al., 2023, 2024), where frequent error types overshadow less common yet pedagogically or computationally significant ones, leading to biased analysis. To assess the balance of a classification taxonomy, we introduce the *Balance Score*. This metric quantifies the evenness of error type distribution using entropy-based uniformity. Given an error type ET_i with proportion computed as:

$$P_i = \frac{|ET_i|}{\sum_{j=1}^{m} |ET_j|}, \tag{4}$$

its entropy contribution is $-P_i \log P_i$ if $|ET_i| > 0$, otherwise 0. The final score is normalized by $\log(m)$, where m is the total number of error types:

$$\text{Balance}(F) = \frac{\sum_{i=1}^{m} -P_i \log P_i}{\log(m)}. \tag{5}$$

A greater Balance value signifies a more uniform distribution, attenuating the long-tail effect and guaranteeing its applicability within educational and computational domains.

3.4 Usability

We argue that a taxonomy with great usability should be understandable for humans and models. So we quantify usability from *model effectiveness* and *human annotation agreement*. Model effectiveness means that LLMs can produce reliable predictions based on a classification taxonomy, thereby enhancing the validity of subsequent error analyses. To quantify this, we evaluate classification performance using *Macro F1* and *Micro F1* scores:

$$\text{Macro\_F1} = \frac{1}{m} \sum_{i=1}^{m} \frac{2 \cdot P(ET_i) \cdot R(ET_i)}{P(ET_i) + R(ET_i)}, \tag{6}$$

$$\text{Micro\_F1} = \frac{2 \cdot P(D) \cdot R(D)}{P(D) + R(D)}, \tag{7}$$

where P and R denote precision and recall, respectively. A high Macro F1 indicates the taxonomy supports stable model performance across both frequent and infrequent error types, mitigating classification bias. A high Micro F1 suggests its robustness in large-scale error detection. The combination of high Macro and Micro F1 scores demonstrates that the taxonomy maintains both category-level consistency and large-scale applicability, ensuring the reliability of error analysis.

For human annotators, the taxonomy should be intuitive and easy to apply, minimizing ambiguity in error categorization. Therefore, we measure inter-annotator consistency in Section 4.3.

4 Experiments

4.1 Experimental Setup

Dataset & Models We use the Cambridge English Write & Improve (W&I) and LOCNESS corpus (Bryant et al., 2019), re-annotated with ERRANT (Bryant et al., 2017). To reduce the uncertainty introduced by multiple grammatical errors in a sentence, we preprocess the dataset by isolating instances with only one error (Fei et al., 2023), compromising 487 instances, as detailed in Appendix B. And we adopt an LLM & human collaborative annotation approach (Appendix D). A detailed description of the dataset is provided in Appendix E. Details of the selected models are provided in Appendix F.

Error Classification Taxonomies We consider four influential error classification taxonomies in error analysis: **POL73** (Politzer and Ramirez, 1973), **TUC74** (Tucker et al., 1974), **BRY17** (Bryant et al., 2017), and **FEI23** (Fei et al., 2023). POL73 and

Model	Exclusivity ($\tau = 0.7$) ↑				Coverage ↑				Balance ↑				Usability (Macro F1 / Micro F1) ↑			
	POL73	TUC74	BRY17	FEI23	POL73	TUC74	BRY17	FEI23	POL73	TUC74	BRY17	FEI23	POL73	TUC74	BRY17	FEI23
Meta-Llama-3-8B-Instruct	0.477	0.772	**0.880**	0.858	0.698	0.160	**0.979**	0.924	0.687	0.210	0.829	**0.878**	0.026 / 0.109	0.005 / 0.006	0.047 / 0.156	**0.112 / 0.357**
Mistral-7B-Instruct-v0.2	0.317	0.095	0.495	**0.840**									0.026 / 0.123	0.001 / 0.004	0.015 / 0.113	**0.085 / 0.310**
Claude-3-Haiku	0.609	0.596	0.742	**0.815**									0.101 / 0.218	0.023 / 0.055	**0.326** / 0.542	0.290 / **0.620**
ChatGPT-4o	0.842	0.703	**0.921**	0.877									0.301 / 0.478	0.061 / 0.099	0.610 / **0.760**	**0.631** / 0.743
Average	0.562	0.542	0.759	**0.848**	0.698	0.160	**0.979**	0.924	0.687	0.210	0.829	**0.878**	0.114 / 0.232	0.023 / 0.041	0.250 / 0.393	**0.280 / 0.508**

Table 1: Performance comparison of classification taxonomies. Higher values indicate better performance in all metrics (Exclusivity, Coverage, Balance, and Usability). The computation of Exclusivity and Usability relies on specific LLMs. Exclusivity is calculated using a confidence threshold of $\tau = 0.7$.

TUC74 are linguistically driven but differ in hierarchical structure and classification logic. BRY17 introduces the rule-based ERRANT toolkit, categorizing errors by part-of-speech and token edit operations. FEI23, grounded in second language acquisition, adopts a cognitive perspective, classifying errors into single-word, inter-word, and discourse-level categories. Appendix G provides a detailed comparison of these taxonomies.

4.2 Main Results

We assess the rationality of different error classification taxonomies in Table 1.

Exclusivity: Ambiguous Type Boundaries Reduce Mutual Exclusivity. Ambiguous definitions reduce mutual exclusivity, leading to annotation inconsistencies. POL73 and TUC74 exhibit high sample overlap, resulting in lower exclusivity than BRY17 and FEI23.

Coverage: A High Number of Categories Does Not Guarantee Comprehensive Coverage. An effective taxonomy must cover both common and rare errors. TUC74, focused on rare domain-specific errors, lacks general grammatical coverage.

Balance: Overgeneralization and Over-Specification Lower Distributional Balance. Excessive granularity creates sparsity, while broad categories cause aggregation.

Usability: Linguistic-Based Taxonomies Reduce Practical Utility. Taxonomies with intricate linguistic distinctions hinder annotation and model integration.

For details, see Appendix H.

4.3 Annotator Agreement Analysis

We measure inter-annotator agreement as a part of Usability, using Cohen's Kappa Index (Cohen, 1960), a standard metric for measuring annotation consistency. The results of three annotators are summarized in Table 2.

BRY17 and FEI23 exhibit higher agreement scores compared to POL73 and TUC74, suggesting that the former taxonomies provide clearer and more well-defined categories. This aligns with the trend observed in Table 1, further substantiating the validity of our evaluation framework. Moreover, the higher agreement in BRY17 and FEI23 reinforces their practical applicability, while the lower scores in POL73 and TUC74 suggest potential ambiguity in category definitions.

Annotator	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

Table 2: Cohen's Kappa Scores for Inter-Annotator Agreement across taxonomies.

To assess the impact of classification granularity, we conducted an ablation study by merging specific error categories and analyzing their effect on key metrics. The results underscore the need for rigorous validation of error classification taxonomies, rather than relying on intuition or convention, highlighting the importance of our proposed evaluation metrics. The detailed results are in Appendix I.

5 Conclusion

We revisit error classification taxonomies in error analysis by proposing a systematic evaluation framework based on exclusivity, coverage, balance, and usability, evaluated on our own annotated dataset. Our experiments validate the effectiveness of these metrics and underscore the need for a systematic assessment of classification taxonomies. Results demonstrate that different taxonomies exhibit trade-offs, with some excelling in exclusivity and coverage while others offer better usability. These findings highlight the importance of well-structured taxonomies for reliable error analysis and annotation consistency.

Limitations

One key limitation of this study is Exclusivity and Usability rely on large language models (LLMs) for computation. While our experimental results demonstrate the reasonableness of these metrics, LLMs inherently introduce biases that may lead to preferences for certain error classification frameworks over others. Future work could explore model-agnostic approaches to mitigate such biases.

Another limitation stems from the dataset scope. Our analysis is based on the W&I+LOCNESS dataset, which, while widely used, may not fully capture the diversity of learner error patterns across different proficiency levels, native languages, and writing contexts. Extending our evaluation to more diverse datasets could improve the generalizability of our findings.

Ethics Statement

We conduct our experiments using the publicly available W&I+LOCNESS dataset, which does not contain sensitive data. All models used are also publicly available, and we have properly cited their sources. The datasets and models are utilized in accordance with their intended purposes. For human agreement evaluation, we employed three postgraduate students specializing in foreign linguistics and applied linguistics as part-time annotators. Each annotator completed the entire annotation process within approximately 20 working hours and was compensated at a rate of $50 per hour.

References

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jana Svoboda Bertkua. 1974. An analysis of english learner speech. *Language Learning*, 24(2):279–286.

Ellen Bialystok, Heidi Dulay, Marina Burt, and Stephen Krashen. 1982. Language two. *The Modern Language Journal*, 67:273.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Marina K. Burt. 1975. Error analysis in the adult efl classroom. *TESOL Quarterly*, 9(1):53–63.

ALICE Y. W. CHAN. 2010. Toward a taxonomy of written errors: Investigation into the written errors of hong kong cantonese esl learners. *TESOL Quarterly*, 44(2):295–319.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Stephen Pit Corder. 1967. The significance of learner's errors.

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 Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Marina Dodigovic. 2007. Artificial intelligence and second language learning: An efficient approach to error remediation. *Language Awareness - LANG AWARE*, 16:99–113.

Heidi C. Dulay and Marina K. Burt. 1972. Goofing: An indicator of children's second language learning strategies. *Language Learning*, 22(2):235–252.

Heidi C. Dulay and Marina K. Burt. 1974. Errors and strategies in child second language acquisition. *TESOL Quarterly*, 8(2):129–136.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.

Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025, Online. Association for Computational Linguistics.

Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*, 1st edition. Routledge.

Carl James. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*, 1st edition. Routledge.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Reza Kafipour and Khojasteh Laleh. 2012. A comparative taxonomy of errors made by iranian undergraduate learners of english. *Canadian Social Science*, 8:18–24.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. 2023. ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 83–99, Toronto, Canada. Association for Computational Linguistics.

Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581. Cambridge University Press Cambridge.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ali Özkayran and Emrullah Yılmaz. 2020. Analysis of higher education students' errors in english writing tasks. *Advances in Language and Literary Studies*.

Robert L. Politzer and Arnulfo G. Ramirez. 1973. An error analysis of the spoken english of mexican-american pupils in a bilingual school and a monolingual school. *Language Learning*, 23(1):39–61.

Annafi'in Nur Rixha, Idrus Alhamid, Siti Rokhmah, and Syamsir Bin Ukka. 2021. Surface strategy taxonomy: Grammatical errors analysis in the third-semester students' descriptive essay. *KARIWARI SMART : Journal of Education Based on Local Wisdom*, 1(2):36–48.

Suhono Suhono. 2017. Surface strategy taxonomy on the efl students composition a study of error analysis. 1:1–30.

G R Tucker, M K Burt, and C Kiparsky. 1974. The gooficon: A repair manual for english. *TESOL Quarterly*, 8(2):191.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *Sociologica*, 18(2):67–85.

Yaredi Waruwu and Afore Tahir Harefa. 2024. An error analysis of communicative effect taxonomy of the tenth grade students' writing of descriptive text. *RETORIKA: Jurnal Ilmu Bahasa*, 10(2):626–636.

Lydia White and Toronto. Bilingual Education Project Ontario Institute for Studies in Education. 1977. Error analysis and error correction in adult learners of english as a second language. working papers on bilingualism, no. 13 [microform]. Accessed: 09 February 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Helen Yannakoudakis, Marek Rei, Øistein E. Andersen, and Zheng Yuan. 2017. Neural sequence-labelling models for grammatical error correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2795–2806, Copenhagen, Denmark. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024. Excgec: A benchmark of edit-wise explainable chinese grammatical error correction. *Preprint*, arXiv:2407.00924.

Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. 2024. A systematic review on long-tailed learning. *Preprint*, arXiv:2408.00483.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. *Preprint*, arXiv:2110.04596.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

A Notation Table

As shown in Table 3, this appendix provides a list of relevant symbols and their meanings for reference.

B Data Preprocessing for Single-Error Isolation

The original dataset contains numerous sentences with n $(n > 1)$ grammatical errors. If such sentences are directly input into the model for error classification, it introduces ambiguity—making it unclear which error the model should focus on. This uncertainty can negatively impact prediction consistency. To mitigate this issue, we preprocess the dataset by isolating individual grammatical errors. Specifically, for each sentence containing n errors, we generate n separate instances, each retaining only one grammatical error while correcting

Symbol	Description
$C^{(x)}$	The placement confidence of sample x
$C_i^{(x)}$	Placement confidence of the i-th predicted error type of sample x in the top-k setting
$\hat{Y}^{(x)}$	Predicted error type of sample x
$\hat{Y}_i^{(x)}$	The i-th predicted error type of sample x in the top-k setting
$\tilde{Y}$	Gold reference of sample x
F	Error classification taxonomy
F_i	The i-th error classification taxonomy
n	Number of error classification taxonomies
ET_j	The j-th error type
m	Number of error types in the corresponding error classification taxonomy F
τ	the predefined confidence threshold
k	the selection parameter in the Top-K Prompting Strategy

Table 3: Notation Table.

the others. This ensures that each model input is unambiguous, leading to more reliable classification results.

C Strategies for Robust Confidence Estimation

To enhance the robustness of *confidence estimation*, we employ three strategies (Xiong et al., 2024) that mitigate overconfidence and improve the reliability of confidence scores:

Top-K Prompting Strategy. One effective approach to reducing overconfidence in LLMs is recognizing the existence of multiple plausible answers. To incorporate this principle, we modify the prompt to explicitly request the generation of the top 3 distinct error types along with their respective confidence scores. By enforcing diversity among the generated error types, this strategy encourages a more calibrated confidence estimation by considering multiple reasonable interpretations.

Self-random Sampling Strategy. Analyzing the variation among multiple responses to the same input can further aid in confidence calibration. This strategy leverages the inherent randomness of LLMs by issuing the same prompt multiple times, thereby introducing controlled variability in the responses. To ensure sufficient diversity in the generated results, we set the temperature to 0.7 across all model configurations.

Average-Confidence (Avg-Conf) Aggregation Strategy. Once multiple responses are generated, an aggregation mechanism is necessary to effectively capture and utilize the observed variability. We adopt the Avg-Conf aggregation strategy, which computes the final confidence distribution by averaging the confidence scores across multiple samples. This approach enables a more robust estimation of confidence levels by accounting for response fluctuations.

By integrating these strategies, we enable LLMs to articulate confidence scores more precisely, leading to improved robustness and trustworthiness in decision-making. For more details, please refer to Xiong et al. (2024).

D LLM & Human Collaborative Annotation Approach Details

In this process, we strictly adhered to the standards and best practices outlined in Törnberg (2024) to ensure that LLM-based annotations are reliable, reproducible, and ethical. First, we utilize the LLM for pre-annotation to reduce the manual workload and improve annotation efficiency. LLM-generated annotations are then reviewed by the authors through an annotation refinement process, where prompts are iteratively refined and optimized to ensure that the LLM fully understands the task requirements. Following the first review, two professional linguists conducted a second round of evaluation. Given the first-round labels, they re-evaluated the data. If discrepancies arose between the linguists and the initial annotation, they engaged in discussions with the authors to finalize the label. This multi-step annotation workflow ensured the reliability and consistency of our dataset.

E Dataset Details

Our dataset consists of 487 samples, each containing a single edit. Each sample is annotated with four labels corresponding to different error classification taxonomies: POL73, TUC74, BRY17, and FEI23. The original dataset does not contain any sensitive personal data, and all data has been anonymized to ensure privacy. Table 4 provides an overview of the dataset statistics.

Attribute	Value
Total Samples	487
Edits per Sample	1
Annotation taxonomies	POL73, TUC74, BRY17, FEI23

Table 4: Statistics of the dataset.

F Model Details

In this study, we employed multiple large language models (LLMs) to evaluate the rationality of different error classification taxonomies in error analysis: **Meta-Llama-3-8B-Instruct** (AI@Meta, 2024), **Mistral-7B-Instruct-v0.2** (Jiang et al., 2023), **Claude-3-Haiku** (Anthropic, 2024), and **ChatGPT-4o** (OpenAI, 2024). These models were selected for their advanced text comprehension and classification capabilities, particularly in handling complex linguistic structures. Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.2 were deployed on a local server equipped with an A800 GPU for inference, while Claude-3-Haiku and ChatGPT-4o were accessed via API calls.

G Error Classification Taxonomy Details

G.1 POL73

This hierarchical error taxonomy, based on a traditional descriptive taxonomy in linguistics, divides errors into three main categories: **Morphology**, **Syntax**, and **Vocabulary**. Each category contains subcategories with definitions and examples for clarity. Figure 1 provides an overview of this taxonomy. Below, we provide a detailed description of this taxonomy:

1 Morphology Errors
Definition: Errors related to the form or structure of words, often involving affixes, tense, agreement, or comparative forms.
1.1 Indefinite Article Incorrect
Definition: Errors in the use of "a" or "an", such as using "a" before vowels or "an" before consonants.

> **Error sentence:** He saw an apple tree with a egg.
> **Correct sentence:** He saw an apple tree with an egg.

1.2 Possessive Case Incorrect
Definition: Errors in forming possessives, including omission, misuse, or incorrect placement of apostrophes.

> **Error sentence:** Mother's Linda came to visit.
> **Correct sentence:** Linda's mother came to visit.

1.3 Third-person Singular Verb Incorrect
Definition: Errors in using third-person singular verb forms, including omission or incorrect addition of "s" or "es."

> **Error sentence:** She walk to school every day.
> **Correct sentence:** She walks to school every day.

1.4 Simple Past Tense Incorrect
Definition: Errors in forming the simple past tense for both regular and irregular verbs.
1.4.1 Regular Past Tense
Definition: Errors involving omission or incorrect addition of "-ed" to regular verbs.

> **Error sentence:** He walk to the store.
> **Correct sentence:** He walked to the store.

1.4.2 Irregular Past Tense
Definition: Errors involving incorrect formation of the past tense for irregular verbs.

> **Error sentence:** He goed to the park.
> **Correct sentence:** He went to the park.

1.5 Past participle incorrect
Definition: Errors in forming or using past participles correctly.

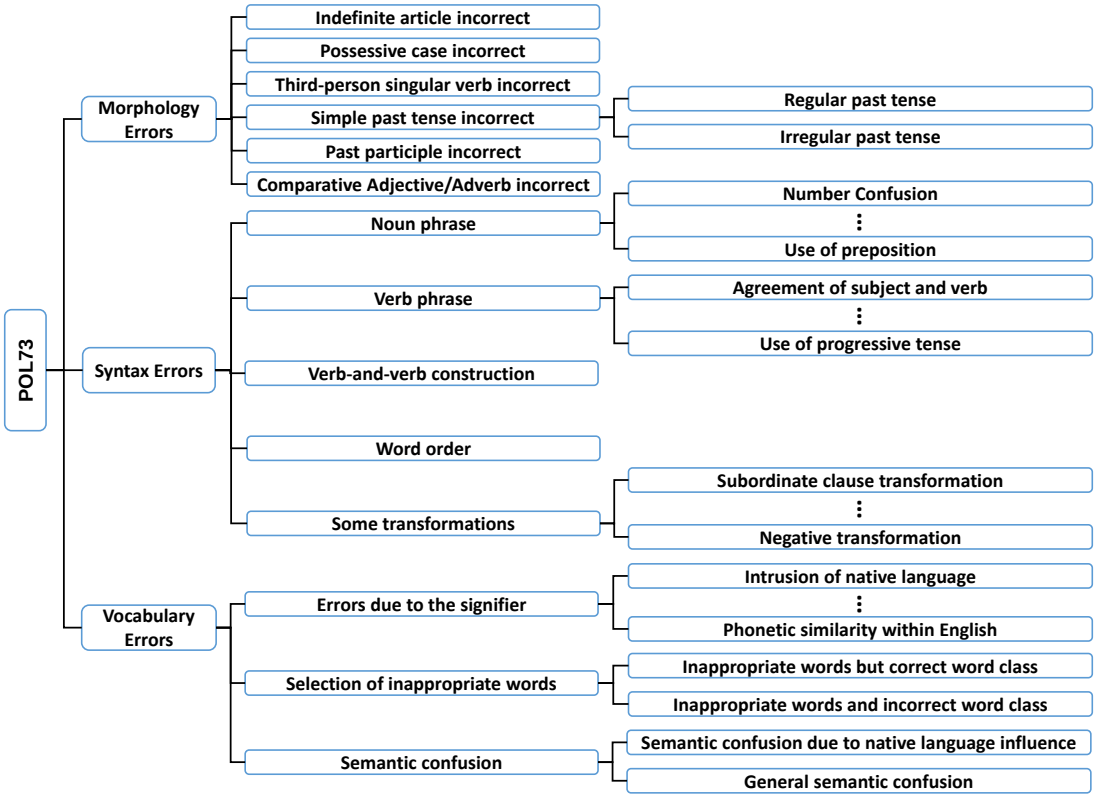

Figure 1: Overview of the POL73 Error Classification Taxonomy. The vertical ellipsis indicates that the category has additional subcategories not fully expanded here.

Error sentence: He was call.
Correct sentence: He was called.

1.6 Comparative Adjective/Adverb incorrect
Definition: Errors in forming or using comparative forms of adjectives and adverbs.

Error sentence: This house is more bigger.
Correct sentence: This house is bigger.

2 Syntax Errors
Definition: Errors related to sentence structure, including issues with phrases, word order, and transformations.

2.1 Noun Phrase
Definition: Errors in constructing or modifying a noun phrase.

2.1.1 Determiners
Definition: Errors in the omission, selection, or placement of determiners.

Error sentence: He bought the apple from a market.
Correct sentence: He bought an apple from a market.

2.1.2 Nominalization
Definition: Errors in using nominalized forms, such as substituting a base verb or incorrect structure for the "-ing" form in appropriate contexts.

Error sentence: He improved the dish by to cook it longer.
Correct sentence: He improved the dish by cooking it longer.

2.1.3 Number Confusion
Definition: Misuse of singular or plural forms.

Error sentence: She has many friend.
Correct sentence: She has many friends.

2.1.4 Use of Pronoun
Definition: Errors in pronoun reference,

form, or usage, including omission, redundancy, substitution, or incorrect agreement with antecedents.

> **Error sentence:** Her went to the store.
> **Correct sentence:** She went to the store.

2.1.5 Use of Preposition
Definition: Errors involving the misuse, omission, or substitution of prepositions in relation to verbs, noun phrases, or expressions of location and direction.

> **Error sentence:** He is good in math.
> **Correct sentence:** He is good at math.

2.2 Verb Phrase
Definition: Errors related to verb use, including tense, form, and omission.
2.2.1 Omission of Verb
Definition: Errors involving the omission of a necessary verb, including main verbs and auxiliary verbs such as forms of "be".

> **Error sentence:** He to school yesterday.
> **Correct sentence:** He went to school yesterday.

2.2.2 Use of Progressive Tense
Definition: Errors in forming or using the progressive tense, including omission of the auxiliary verb "be", incorrect use of the "-ing" form, or substitution of the progressive where it is not expected.

> **Error sentence:** He is play football.
> **Correct sentence:** He is playing football.

2.2.3 Agreement of Subject and Verb
Definition: Errors where the subject and verb fail to agree in number, person, or tense.

> **Error sentence:** I didn't know what it is.
> **Correct sentence:** I didn't know what it was.

2.3 Verb-and-Verb Construction
Definition: Errors in embedding one verb phrase into another, often involving omissions or misuse of linking elements like "to" or incorrect verb forms.

> **Error sentence:** She wants goes home.
> **Correct sentence:** She wants to go home.

2.4 Word Order
Definition: Errors in the logical arrangement of words in a sentence, including incorrect placement of subjects, verbs, objects, or modifiers.

> **Error sentence:** To school she goes every day.
> **Correct sentence:** She goes to school every day.

2.5 Some Transformations
Definition: Errors in applying specific grammatical transformations, such as passive voice, negations, questions, or subordinate clauses.
2.5.1 Passive Transformation
Definition: Errors in the formation or structure of passive voice sentences.

> **Error sentence:** The bird got saved.
> **Correct sentence:** The bird was saved.

2.5.2 Negative Transformation
Definition: Errors in constructing negative sentences, including misuse of auxiliary verbs or double negations.

> **Error sentence:** She doesn't likes the movie.
> **Correct sentence:** She doesn't like the movie.

2.5.3 Question Transformation
Definition: Errors in forming questions, often involving incorrect word order or auxiliary usage.

> **Error sentence:** Where he is going?
> **Correct sentence:** Where is he going?

2.5.4 There Transformation
Definition: Errors in using "there" as a subject, including misuse or omission in existential constructions.

> **Error sentence:** There is many people in the room.
> **Correct sentence:** There are many people in the room.

2.5.5 Subordinate Clause Transformation
Definition: Errors in constructing subordinate clauses, such as unnecessary elements or incorrect conjunction usage.

> **Error sentence:** I know that where she lives.
> **Correct sentence:** I know where she lives.

3 Vocabulary Errors
Definition: Errors caused by confusion in word meaning, selection, or form, often due to similar sounds, inappropriate use in grammatical constructions, or meaning substitution.
3.1 Errors Due to the Signifier
Definition: The form or sound of a word leads to a misinterpretation or incorrect usage in a new linguistic context.
3.1.1 Intrusion of Native Language
Definition: Errors resulting from the direct influence of the speaker's native language on English usage.

> **Error sentence:** He no lo mat.
> **Correct sentence:** He does not kill it

3.1.2 Phonetic Similarities Between Native Language and English
Definition: Errors caused by words resembling native language phonetics but incorrect in English.

> **Error sentence:** The parablem was difficult to solve.
> **Correct sentence:** The problem was difficult to solve.

3.1.3 Phonetic Similarity within English
Definition: Errors stemming from mishearing or phonetic misinterpretation of English words.

> **Error sentence:** He tried to kid the mosquito.
> **Correct sentence:** He tried to kill the mosquito.

3.1.4 New Creations
Definition: Errors caused by inventing nonexistent English words for unclear reasons.

> **Error sentence:** He was drownding in the water.
> **Correct sentence:** He was drowning in the water.

3.2 Selection of Inappropriate Words
Definition: Errors where a word with an incorrect meaning or grammatical role is chosen, despite the sentence being grammatically constructed correctly.
3.2.1 Inappropriate Words But Correct Word Class
Definition: Errors where the word belongs to the correct grammatical class but is semantically inappropriate.

> **Error sentence:** They won't help together.
> **Correct sentence:** They won't help each other.

3.2.2 Inappropriate Words and Incorrect Word Class
Definition: Errors where the word is both grammatically and semantically inappropriate.

> **Error sentence:** The ant has an open in the ground.
> **Correct sentence:** The ant has a hole in the ground.

3.3 Semantic Confusion
Definition: Errors occur when a word or phrase is used incorrectly due to misunderstanding its meaning or the influence of another language.
3.3.1 Semantic Confusion Due to Native Language Influence
Definition: Errors caused by incorrect meaning or usage influenced by the native language.

Error sentence: The little ant was to the house. [note: (Spanish, fue) == (English, was) == (English, were)]
Correct sentence: The little ant went to the house.

3.3.2 General Semantic Confusion
Definition: Errors caused by inappropriate word usage that cannot be traced to native language influence.

Error sentence: The man came into the water.
Correct sentence: The man fell into the water.

G.2 TUC74

This error taxonomy is a hierarchical taxonomy based on the structural, functional, and grammatical characteristics of English, encompassing errors in syntax, morphology, and semantic alignment within sentence construction. Each category contains subcategories with definitions and examples for clarity. Figure 2 provides an overview of this taxonomy. Below, we provide a detailed description of this taxonomy:

1 The Skeleton of English Clauses
Definition: Fundamental grammatical errors that compromise the basic structure of English clauses, including omissions and misplacements of critical components.
1.1 Missing Parts
Definition: Errors involving the omission of essential sentence components necessary for grammatical completeness and structural clarity.
1.1.1 Surrogate Subject Missing: there and it
Definition: Missing placeholder subjects "there" or "it," which are required in English sentence structure.

Error sentence: Was a riot last night.
Correct sentence: There was a riot last night.

1.1.2 Simple Predicate Missing: *be*
Definition: Missing the simple predicate "be," which is required when the predicate consists of adjectives or noun phrases.

Error sentence: My sisters very pretty.
Correct sentence: My sisters are very pretty.

1.1.3 Object Pronoun Missing
Definition: Missing object pronouns (direct or indirect) in places where verbs require them.

Error sentence: Donald is mean so no one likes.
Correct sentence: Donald is mean so no one likes him.

1.1.4 Subject Pronoun Missing
Definition: Missing subject pronouns where they are required, especially after subordinate conjunctions.

Error sentence: He worked until fell over.
Correct sentence: He worked until he fell over.

1.2 Misordered Parts
Definition: Errors involving incorrect word order that disrupt the standard syntactic structure of English sentences.
1.2.1 Verb Before Subject
Definition: Incorrect word order where the verb precedes the subject in declarative sentences.

Error sentence: Escaped the professor from prison.
Correct sentence: The professor escaped from prison.

1.2.2 Subject and Object Permuted
Definition: Errors occur when subject and object positions are reversed in declarative sentences.

Error sentence: English use many countries.
Correct sentence: Many countries use English.

2 The Auxiliary System
Definition: Errors in the misuse, omission,

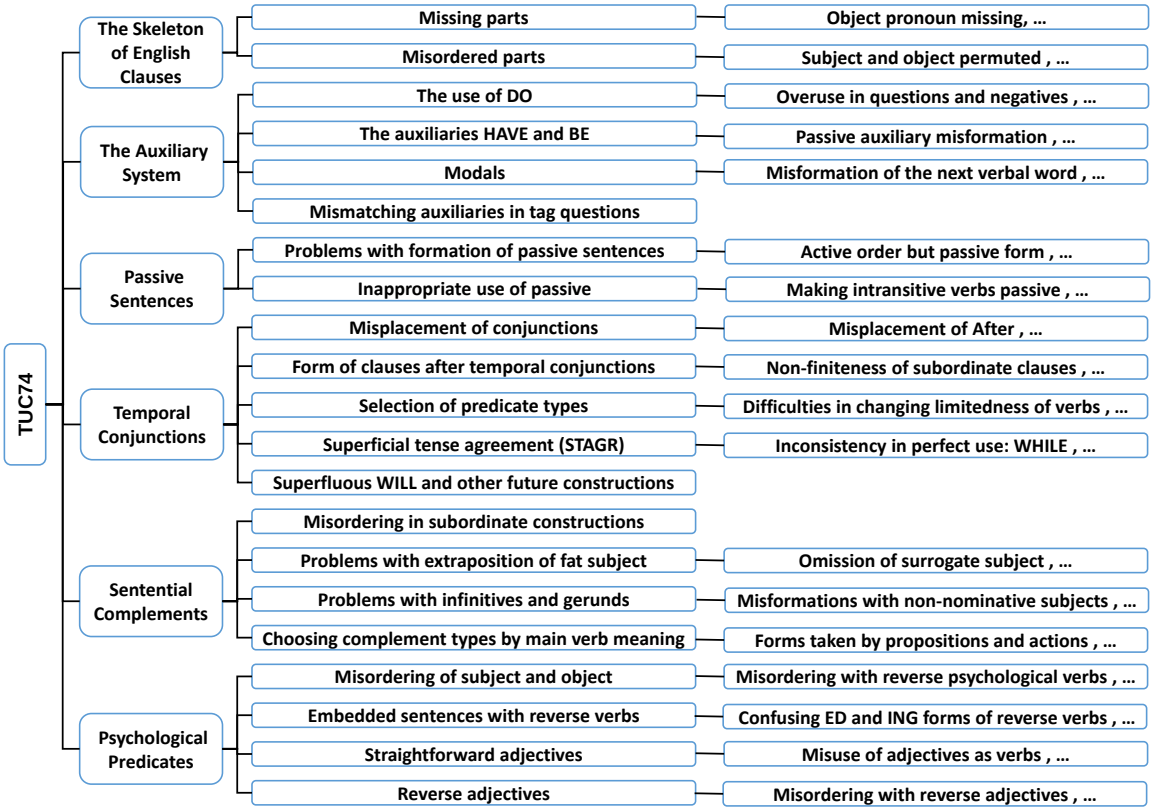

Figure 2: Overview of the TUC74 Error Classification Taxonomy. The horizontal ellipsis indicates that the category has additional subcategories not fully expanded here.

or misalignment of auxiliary verbs ("do," "have," "be," modals) in questions, negatives, affirmatives, tense structures, or tag questions.

2.1 The Use of *DO*

Definition: Errors related to the auxiliary verb "do," including its overuse, underuse, or misuse in forming questions, negatives, affirmatives, or maintaining consistent tense structures in English clauses.

2.1.1 Overuse in Questions and Negatives

Definition: Incorrect use of "do" in questions or negatives, where it appears with modal auxiliaries (can, must, should, etc.) or other auxiliaries (have, be) that already fulfill the grammatical function.

> **Error sentence:** Does she have come yet?
> **Correct sentence:** Has she come yet?

2.1.2 Underuse in Questions

Definition: Failure to include do in questions when there are no auxiliary verbs, leading to incorrect word order or ungrammatical structures.

> **Error sentence:** Why we bow to each other?
> **Correct sentence:** Why do we bow to each other?

2.1.3 Overuse in Affirmative Sentences

Definition: "Do" appears in a clause if there is no auxiliary, and the clause is a question or a negative. It does not appear in affirmative sentences.

> **Error sentence:** I did leave yesterday.
> **Correct sentence:** I left yesterday.

2.1.4 DO Missing from Negatives

Definition: The auxiliary verb "do" is missing in negative sentences without other auxiliaries. In such cases, "do" must appear, and "not" must follow it to form the correct structure.

Error sentence: I practice not religion.
Correct sentence: I do not practice religion.

2.1.5 Tense Misplacement
Definition: Errors involving placing tense on more than one verbal word or using conflicting tenses in a clause.

Error sentence: Do you saw her already?
Correct sentence: Did you see her already?

2.2 The Auxiliaries *HAVE* and *BE*
Definition: Errors involving the misuse, omission, or incorrect combination of "have" and "be" in forming perfect, progressive, passive constructions, or linking non-verb predicates.
2.2.1 Misformation of Perfect and Progressive Aspects
Definition: Errors involving the incorrect formation of either the perfect aspect (combination of "have" and past participle) or the progressive aspect (combination of "be" and present participle).

Error sentence: I have saw Broadway.
Correct sentence: I have seen Broadway.

2.2.2 Passive Auxiliary Misformation
Definition: Errors involving the omission or misuse of "be" in forming the passive voice.

Error sentence: I have impressed with Plato.
Correct sentence: I am impressed with Plato.

2.2.3 BE Missing
Definition: Errors caused by omitting "be" before non-verb predicates (e.g., adjectives, nouns, adverbs, etc.), specifically in sentences that are not progressive or passive constructions.

Error sentence: The bus always full of people.
Correct sentence: The bus is always full of people.

2.2.4 DO Misused with BE
Definition: Errors caused by incorrectly using "do" before "be" in questions or negatives.

Error sentence: Do they be happy?
Correct sentence: Are they happy?

2.3 Modals
Definition: Errors in the use of modal verbs, affecting their grammatical role or the structure of the sentence.
2.3.1 Misformation of the Next Verbal Word
Definition: Errors involving the use of affixed forms (e.g., -ing, -ed) or the addition of "to" after modal verbs, which require a base form of the verb.

Error sentence: I can going if you can.
Correct sentence: I can go if you can.

2.3.2 Misunderstanding of Tense with Modals
Definition: Errors stemming from a misunderstanding of modal tense properties.

Error sentence: I must can catch this train.
Correct sentence: I must catch this train.

2.4 Mismatching Auxiliaries in Tag Questions
Definition: Errors caused by using a different auxiliary in the tag question than the one used in the main clause.

Error sentence: She has been smoking less, isn't it?
Correct sentence: She has been smoking less, hasn't she?

3 Passive Sentences
Definition: Errors in passive voice construction or usage, including verb form issues, structural mismatches, preposition errors, or inappropriate application in certain contexts.

3.1 Problems with Formation of Passive Sentences
Definition: Errors in passive sentence construction, including misuse of "be," mismatched subject-verb relations, incorrect prepositions, or confusion between active and passive forms.

3.1.1 Misformation of Passive Verb
Definition: Errors caused by omitting or misplacing the auxiliary "be" in passive constructions, or failing to use the past participle form of the verb.

> **Error sentence:** The bread finished.
> **Correct sentence:** The bread is finished.

3.1.2 Active Order but Passive Form
Definition: Errors caused by retaining the active sentence order while incorrectly using the passive form, resulting in a mismatch between the subject and the sentence structure.

> **Error sentence:** The government was forbidden the people to grow opium.
> **Correct sentence:** The people were forbidden to grow opium by the government.

3.1.3 Absent or Wrong Preposition Before Agent
Definition: Errors caused by omitting or using the wrong preposition (e.g., "by") before the agent in a passive sentence.

> **Error sentence:** My brother was held up the traffic jam.
> **Correct sentence:** My brother was held up by the traffic jam.

3.1.4 Passive Order but Active Form
Definition: Errors caused by using active verbs while following the word order of passive sentences, leading to mismatched sentence structure.

> **Error sentence:** English use many countries.
> **Correct sentence:** English is used by many countries.

3.2 Inappropriate Use of Passive
Definition: Errors involving the misuse of passive voice, such as applying it to intransitive verbs or overextending it in complex sentences, leading to grammatical or logical inconsistencies.

3.2.1 Making Intransitive Verbs Passive
Definition: Errors caused by attempting to form the passive voice using intransitive verbs, which lack an object to become the subject in a passive construction.

> **Error sentence:** He was arrived early.
> **Correct sentence:** He arrived early.

3.2.2 Misusing Passives in Complex Sentences
Definition: Errors caused by incorrectly applying passive constructions across multiple clauses in a complex sentence, leading to confusion about the subject and clause relations.

> **Error sentence:** I was suggested by Mrs. Sena to forget about this project.
> **Correct sentence:** Mrs. Sena suggested that I forget about this project.

4. Temporal Conjunctions
Definition: Errors in the use, placement, or grammatical constructions associated with temporal conjunctions, including issues with clause structure, predicate types, tense agreement, and unnecessary future markers.

4.1 Misplacement of Conjunctions
Definition: Errors caused by incorrectly positioning conjunctions like "after," "since," or "while," leading to confusion about the sequence, causality, or timing of events in a sentence.

4.1.1 Misplacement of After
Definition: Errors involving incorrect placement or usage of "after" when describing the sequence of two events.

> **Error sentence:** I got up after I brushed my teeth.

Correct sentence: After I got up, I brushed my teeth.

4.1.2 Misplacement of Since

Definition: Errors involving incorrect placement or usage of "since" when describing the temporal relationship between two events.

Error sentence: He broke his leg since he has thrown away his skis.
Correct sentence: Since he broke his leg, he has thrown away his skis.

4.1.3 Misplacement of While

Definition: Errors involving incorrect placement or usage of "while" to describe overlapping or interrupting events.

Error sentence: While Getachew knocked on the door, I was doing my homework.
Correct sentence: Getachew knocked on the door while I was doing my homework.

4.2 Form of Clauses After Temporal Conjunctions

Definition: Errors in the structure of subordinate clauses following temporal conjunctions.

4.2.1 Non-Finiteness of Subordinate Clauses

Definition: Errors caused by omitting subjects or using incorrect verb forms in subordinate clauses, making them non-finite when they should be finite.

Error sentence: After him goes, we will read a story.
Correct sentence: After he goes, we will read a story.

4.2.2 Superfluous THAT

Definition: Errors caused by inserting an unnecessary "that" immediately after a subordinate conjunction, disrupting sentence structure.

Error sentence: Since that he has seen her, he has been cheerful.
Correct sentence: Since he has seen her, he has been cheerful.

4.3 Selection of Predicate Types

Definition: Errors in choosing appropriate predicate forms with temporal conjunctions.

4.3.1 Confusion in Unlimited and Limited Verb Selection

Definition: Errors caused by incorrect use of limited or unlimited verbs with temporal conjunctions such as "until" or "after".

Error sentence: He got rich until he married.
Correct sentence: He was rich until he married.

4.3.2 Difficulties in Changing Limitedness of Verbs

Definition: Errors caused by using inherently limited verbs in constructions that require unlimited verbs.

Error sentence: Since the child recovered from measles, he grew well.
Correct sentence: Since the child recovered from measles, he has been growing well.

4.3.3 Misuse of Negatives with Temporal Conjunctions

Definition: Errors involving the use of negatives with temporal conjunctions, where the clause requires a specific form of limited or unlimited verb.

Error sentence: I did it while they didn't look.
Correct sentence: I did it while they weren't looking.

4.3.4 Misuse of End-of-the-Road Predicates

Definition: Errors arising from the use of predicates that denote final or end states (e.g., dead, rotten, finished) in subordinate clauses with temporal conjunctions, where such predicates conflict with the expected progression or transition of states.

Error sentence: The fruit had become rotten until we could eat it.
Correct sentence: The fruit had become rotten before we could eat it.

4.4 Superficial Tense Agreement (STAGR)
Definition: Errors in tense consistency between clauses linked by temporal conjunctions.

4.4.1 Failure to Apply STAGR with BEFORE, AFTER, UNTIL, WHILE, WHEN
Definition: Errors caused by mismatched tenses in clauses connected by temporal conjunctions. This typically occurs when the first verb in each clause does not agree in tense.

> **Error sentence:** When you were here yesterday you have promised to send your picture.
> **Correct sentence:** When you were here yesterday you promised to send your picture.

4.4.2 Inconsistency in Perfect Use: WHILE
Definition: Errors occurring when one clause in a while construction uses the perfect aspect, but the other clause does not. Both clauses must either use the perfect aspect or omit it.

> **Error sentence:** While you have worked, I make phone calls.
> **Correct sentence:** While you worked, I made a phone call.

4.4.3 STAGR Misapplied: SINCE
Definition: Errors caused by applying the STAGR rule (matching tenses) to since clauses, where since requires specific tense relationships between clauses.

> **Error sentence:** They are studying in this school since they are six years old.
> **Correct sentence:** They have been studying in this school since they were six years old.

4.5 Superfluous WILL and Other Future Constructions
Definition: Errors involving unnecessary use of *will* or other future constructions in subordinate clauses when the main clause already indicates future tense.

> **Error sentence:** We will eat after we will pray.
> **Correct sentence:** We will eat after we pray.

5. Sentential Complements
Definition: Errors in subordinate clauses or complements, including word order, surrogate subjects, infinitives, gerunds, or complement forms required by the main verb.

5.1 Misordering in Subordinate Constructions
Definition: Errors caused by incorrect word order in subordinate constructions, where clauses deviate from the subject-verb-object (SVO) pattern or verbs are misplaced.

> **Error sentence:** He says that he no money has.
> **Correct sentence:** He says that he has no money.

5.2 Problems with Extraposition of Fat Subject
Definition: Errors involving the omission or misuse of surrogate subjects like "it" or "there" when a heavy subject, such as a that-clause, is moved to the end of the sentence.

5.2.1 Omission of Surrogate Subject
Definition: Missing the word "it" as the subject when a heavy subject, such as a that-clause, is moved to the end of the sentence.

> **Error sentence:** Is very hard for me to learn English right.
> **Correct sentence:** It is very hard for me to learn English right.

5.2.2 Wrong Surrogate Subject: IT and THERE
Definition: Errors caused by using incorrect surrogate subjects, such as he or that, instead of it or there, in sentences describing the weather, ambient conditions, or with extra posed subjects.

> **Error sentence:** That is funny to see him today.
> **Correct sentence:** It is funny to see him today.

5.3 Problems with Infinitives and Gerunds

Definition: Errors in the use of infinitives and gerunds.

5.3.1 Leaving Out the Subject

Definition: Errors occur when the subject of a subordinate clause is omitted incorrectly, particularly when the subject is not repetitive or differs from the main clause subject.

> **Error sentence:** I think to have my I.D. card in here.
> **Correct sentence:** I think I have my I.D. card in here.

5.3.2 Misformations with Non-Nominative Subjects

Definition: Errors occur when a non-nominative subject is incorrectly used with an infinitive or a gerund, instead of the correct form.

> **Error sentence:** Him to be so rich is unfair.
> **Correct sentence:** For him to be so rich is unfair.

5.3.3 Misformations Without Subjects

Definition: Errors occur when "for" is mistakenly used before an infinitive that lacks a subject, instead of simply using "to" before the verb.

> **Error sentence:** It is necessary for finish the work.
> **Correct sentence:** It is necessary to finish the work.

5.3.4 Special Problems with MAKE, LET, HAVE, FIND

Definition: Errors occur when using make, let, have, or find with infinitives or complements, such as unnecessarily adding "to" or omitting the required "it" as a surrogate subject.

> **Error sentence:** The vacuum cleaner makes easy to clean the house.
> **Correct sentence:** The vacuum cleaner makes it easy to clean the house.

5.3.5 Snatched Subject as Subject of Main Clause

Definition: Errors occur when subject snatching is applied to adjectives or predicates that do not allow it, resulting in ungrammatical constructions.

> **Error sentence:** The President is impossible to be reelected.
> **Correct sentence:** It is impossible that the President will be reelected.

5.3.6 Snatched Subject as Object of Main Clause

Definition: Errors occur when a verb that does not permit subject snatching is used incorrectly to create a construction where the subordinate subject becomes the main clause's object.

> **Error sentence:** A girl was decided to play the piano.
> **Correct sentence:** It was decided that a girl would play the piano.

5.3.7 Misformation of Gerunds After Prepositions

Definition: Errors occur when the complement following a preposition is not in the required gerund or nominal form.

> **Error sentence:** We look forward to see you again.
> **Correct sentence:** We look forward to seeing you again.

5.4 Choosing Complement Types by Main Verb Meaning

Definition: Errors in selecting the correct complement form (e.g., that-clauses, infinitives, or gerunds) based on the meaning and requirements of the main verb.

5.4.1 Forms Taken by Propositions and Actions

Definition: Errors occur when the incorrect complement form is used after verbs. Verbs like think or believe require a that-clause, while verbs like want or stop typically take gerunds or infinitives.

> **Error sentence:** Mark thinks the beans needing fertilizer.
> **Correct sentence:** Mark thinks that the beans need fertilizer.

5.4.2 Difficulty with Verbs Which Select Infinitives
Definition: Errors occur when infinitives are required but incorrectly formed.

> **Error sentence:** I don't expect seeing him today.
> **Correct sentence:** I don't expect to see him today.

5.4.3 Difficulty with Verbs Which Select Gerunds
Definition: Errors occur when gerunds are required but incorrectly formed.

> **Error sentence:** Don't you remember to see her yesterday?
> **Correct sentence:** Don't you remember seeing her yesterday?

5.4.4 Confusion Over Complement Form After Auxiliaries
Definition: Errors occur when the presence of auxiliaries leads to incorrect complement forms, as auxiliaries do not change the selection of complement forms determined by the main verb.

> **Error sentence:** I will enjoy to swim.
> **Correct sentence:** I will enjoy swimming.

6. Psychological Predicates
Definition: Errors in constructing sentences with psychological predicates.
6.1 Misordering of Subject and Object
Definition: Errors caused by incorrect word order between subject and object, particularly with psychological verbs, leading to confusion in the roles of experiencer and stimulus.
6.1.1 Misordering with Reverse Psychological Verbs
Definition: Errors occur when the required stimulus-verb-experiencer word order for reverse psychological verbs is not followed, leading to confusion in meaning.

> **Error sentence:** The cat is on the dinner table, but my father doesn't bother that.
> **Correct sentence:** The cat is on the dinner table, but that doesn't bother my father.

6.1.2 Misordering with Straightforward Psychological Verbs
Definition: Errors occur when experiencer and stimulus roles are reversed for straightforward psychological verbs, which require experiencer-verb-stimulus order.

> **Error sentence:** The party enjoyed Aziz.
> **Correct sentence:** Aziz enjoyed the party.

6.2 Embedded Sentences with Reverse Verbs
Definition: Errors in using reverse psychological verbs, including misused forms, incorrect prepositions, missing elements, or improper handling of complex stimuli.
6.2.1 Using the Experiencer as Subject
Definition: Errors occur when the experiencer is used as the subject of reverse verbs without converting the verb to its participial form with "ED" and adding "be."

> **Error sentence:** I surprise that he likes it.
> **Correct sentence:** I am surprised that he likes it.

6.2.2 Wrong Use of Prepositions with ED Forms
Definition: Errors occur when incorrect prepositions are used after reverse verbs in their ED form, as each verb typically requires specific prepositions.

> **Error sentence:** We were all bored about his teaching.
> **Correct sentence:** We were all bored by his teaching.

6.2.3 Confusing ED and ING Forms of Reverse Verbs
Definition: Errors occur when ING forms (which describe stimuli) are used instead of ED forms (which describe experiencers), or vice versa.

> **Error sentence:** I was surprising that he came.
> **Correct sentence:** I was surprised that he came.

6.2.4 Leaving Out Stimulus or Experiencer
Definition: Errors occur when reverse psychological verbs omit either the stimulus or the experiencer, leading to incomplete or unclear sentences.

> **Error sentence:** Don't go to that movie. It bores.
> **Correct sentence:** Don't go to that movie. It's boring.

6.2.5 Mismanaged Extraposition
Definition: Errors occur when reverse verbs with complex or long stimuli are misstructured, often leading to awkward sentences. These can be corrected by reordering the stimulus or using an extraposition with "it."

> **Error sentence:** Everyone delights that you won the lottery.
> **Correct sentence:** Everyone is delighted that you won the lottery.

6.3 Straightforward Adjectives
Definition: Errors involving the misuse of straightforward adjectives, including treating them as reverse adjectives or misusing them as verbs instead of linking to appropriate verbs.
6.3.1 Misordering with Straightforward Adjectives
Definition: Errors occur when straightforward adjectives are treated as reverse adjectives, with the stimulus used as the subject instead of the experiencer.

> **Error sentence:** It was impatient to me to find out my grade.
> **Correct sentence:** I was impatient to find out my grade.

6.3.2 Misuse of Adjectives as Verbs
Definition: Errors occur when adjectives are misused as verbs instead of appearing after linking verbs like "be," "become," or others such as "feel" or "seem."

> **Error sentence:** Kwame sorries so much that your wife is sick.
> **Correct sentence:** Kwame is sorry so much that your wife is sick.

6.4 Reverse Adjectives
Definition: Errors in the use of reverse adjectives, including misplacing the stimulus and experiencer, misordering in embedded sentences, or mishandling causation with impersonal adjectives.
6.4.1 Misordering with Reverse Adjectives
Definition: Errors occur when reverse adjectives are misused as straightforward adjectives, placing the experiencer as the subject instead of the stimulus.

> **Error sentence:** I am hard to get anything done.
> **Correct sentence:** It is hard for me to get anything done.

6.4.2 Misordering in Embedded Sentences
Definition: Errors occur when reverse adjectives in embedded sentences place the experiencer as the subject instead of the stimulus.

> **Error sentence:** He admits me hard to learn quickly.
> **Correct sentence:** He admits (that) it is hard for me to learn quickly.

6.4.3 Difficulties with Causation
Definition: Errors occur when causation is expressed incorrectly, often by misplacing objects or failing to use "it" with impersonal adjectives like "impossible" or "easy."

> **Error sentence:** The kids make impossible me to work.
> **Correct sentence:** The kids make it impossible for me to work.

G.3 BRY17

This error taxonomy is based on the Part-of-Speech (PoS) tags and the granularity level of the error. Any error type in the taxonomy is prefixed with 'M:', 'R:', or 'U:', depending on whether it describes a **Missing**, **Replacement**, or **Unnecessary**

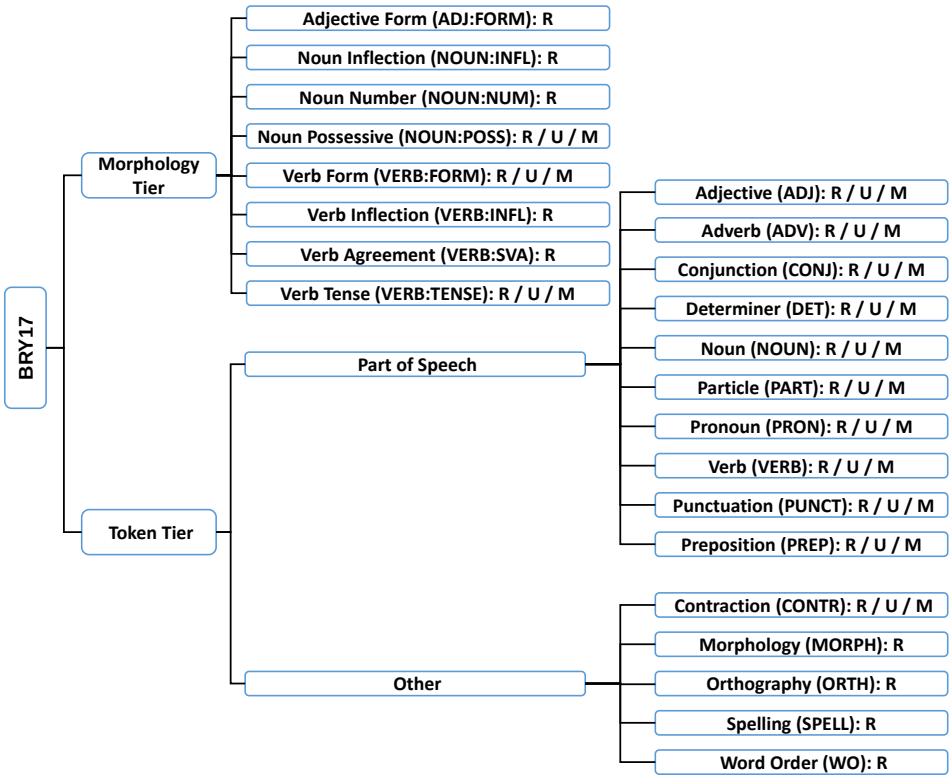

Figure 3: Overview of the BRY17 Error Classification Taxonomy

edit respectively. Each category contains subcategories with definitions and examples for clarity. Figure 3 provides an overview of this taxonomy. Below, we provide a detailed description of this taxonomy:

1 Morphology Tier
Definition: Errors related to morphological structures such as inflection, agreement, and verb forms.

1.1 Adjective Form (ADJ:FORM) [R]
Definition: Errors in the form of adjectives, such as incorrect comparative or superlative forms.

> **Error sentence:** This house is more bigger than the other.
> **Correct sentence:** This house is bigger than the other.

1.2 Noun Inflection (NOUN:INFL) [R]
Definition: Count-mass noun errors.

> **Error sentence:** His advices were not helpful.
> **Correct sentence:** His advice was not helpful.

1.3 Noun Number (NOUN:NUM) [R]
Definition: Errors in the singular or plural form of nouns.

> **Error sentence:** She has many friend.
> **Correct sentence:** She has many friends.

1.4 Noun Possessive (NOUN:POSS) [R/U/M]
Definition: Errors in the possessive case, including incorrect or missing apostrophes, or unnecessary possessive markers.

> **Error sentence:** This is Johns book.
> **Correct sentence:** This is John's book.

1.5 Verb Form (VERB:FORM) [R/U/M]
Definition: Errors in verb form, such as misuse of gerunds, infinitives, or participles.

> **Error sentence:** He improved the dish by to cook it longer.
> **Correct sentence:** He improved the dish by cooking it longer.

1.6 Verb Inflection (VERB:INFL) [R]
Definition: Misapplication of tense morphology.

> **Error sentence:** She getted a new car yesterday.
> **Correct sentence:** She got a new car yesterday.

1.7 Verb Agreement (VERB:SVA) [R]
Definition: Subject-verb agreement errors, such as mismatched number or person.

> **Error sentence:** They was happy.
> **Correct sentence:** They were happy.

1.8 Verb Tense (VERB:TENSE) [R/U/M]
Definition: Wrong choice of inflectional and periphrastic tense, modal verbs, and passivization.

> **Error sentence:** He has went to the park yesterday.
> **Correct sentence:** He went to the park yesterday.

2 Token Tier
Definition: Errors related to individual tokens, including words and punctuation.
2.1 Part of Speech
2.1.1 Adjective (ADJ) [R/U/M]
Definition: Errors in the use, omission, or addition of adjectives.

> **Error sentence:** She is a beauty dancer.
> **Correct sentence:** She is a beautiful dancer.

2.1.2 Adverb (ADV) [R/U/M]
Definition: Errors in the use, omission, or addition of adverbs.

> **Error sentence:** She sings good.
> **Correct sentence:** She sings well.

2.1.3 Conjunction (CONJ) [R/U/M]
Definition: Errors in the use, omission, or addition of conjunctions.

> **Error sentence:** She likes apples oranges.
> **Correct sentence:** She likes apples and oranges.

2.1.4 Determiner (DET) [R/U/M]
Definition: Errors in the use, omission, or addition of determiners.

> **Error sentence:** She bought the a apple.
> **Correct sentence:** She bought the apple.

2.1.5 Noun (NOUN) [R/U/M]
Definition: Errors in the use, omission, or addition of nouns.

> **Error sentence:** She brought her friend friend.
> **Correct sentence:** She brought her friend.

2.1.6 Particle (PART) [R/U/M]
Definition: Errors involving particles, such as their incorrect placement, omission, or addition.

> **Error sentence:** He got up into the car.
> **Correct sentence:** He got into the car.

2.1.7 Pronoun (PRON) [R/U/M]
Definition: Errors involving pronouns, such as incorrect form, extra pronouns, or omitted pronouns.

> **Error sentence:** She herself went to the store herself.
> **Correct sentence:** She went to the store herself.

2.1.8 Verb (VERB) [R/U/M]
Definition: Errors involving the incorrect verb form, addition of redundant verbs, or omission of verbs.

> **Error sentence:** He to school every day.
> **Correct sentence:** He goes to school every day.

2.1.9 Punctuation (PUNCT) [R/U/M]
Definition: Errors in punctuation, including incorrect use, extra punctuation marks, or missing punctuation.

> **Error sentence:** She said, "Hello".
> **Correct sentence:** She said, "Hello."

2.1.10 Preposition (PREP) [R/U/M]
Definition: Errors in the use, omission, or addition of prepositions.

> **Error sentence:** He is good math.
> **Correct sentence:** He is good at math.

2.2 Other
2.2.1 Contraction (CONTR) [R/U/M]
Definition: Errors in the use, omission, or addition of contractions.

> **Error sentence:** I isn't ready for the exam.
> **Correct sentence:** I'm not ready for the exam.

2.2.2 Morphology (MORPH) [R]
Definition: Errors where tokens share the same lemma but differ in other grammatical or syntactical attributes.

> **Error sentence:** The boy runned fast.
> **Correct sentence:** The boy ran fast.

2.2.3 Orthography (ORTH) [R]
Definition: Errors in capitalization, whitespace, or other orthographic conventions.

> **Error sentence:** he went to School.
> **Correct sentence:** He went to school.

2.2.4 Spelling (SPELL) [R]
Definition: Errors in the spelling of words.

> **Error sentence:** She recieved the package.
> **Correct sentence:** She received the package.

2.2.5 Word Order (WO) [R]
Definition: Errors in the arrangement of words within a sentence.

> **Error sentence:** To school she goes every day.
> **Correct sentence:** She goes to school every day.

G.4 FEI23

This taxonomy classifies grammatical errors into three levels based on cognitive complexity (**Single-word level**, **Inter-word level**, and **Discourse-level**) addressing different aspects of language learning and comprehension. Each category contains sub-categories with definitions and examples for clarity. Figure 4 provides an overview of this taxonomy. Below, we provide a detailed description of this taxonomy:

1 Single-word Level Error
Definition: Errors confined to a single word, typically involving spelling, contractions, or capitalization.
1.1 Spelling Error
Definition: Incorrect spelling of a word.

> **Error sentence:** She recieved the letter.
> **Correct sentence:** She received the letter.

1.2 Contraction Error
Definition: Incorrect use or formation of contractions.

> **Error sentence:** I isn't ready.
> **Correct sentence:** I'm not ready.

1.3 Orthography Error
Definition: Errors in capitalization or other writing conventions.

> **Error sentence:** he went to school.
> **Correct sentence:** He went to school.

2 Inter-word Level Error
Definition: Errors involving relationships between multiple words, including grammar, morphology, and word usage.

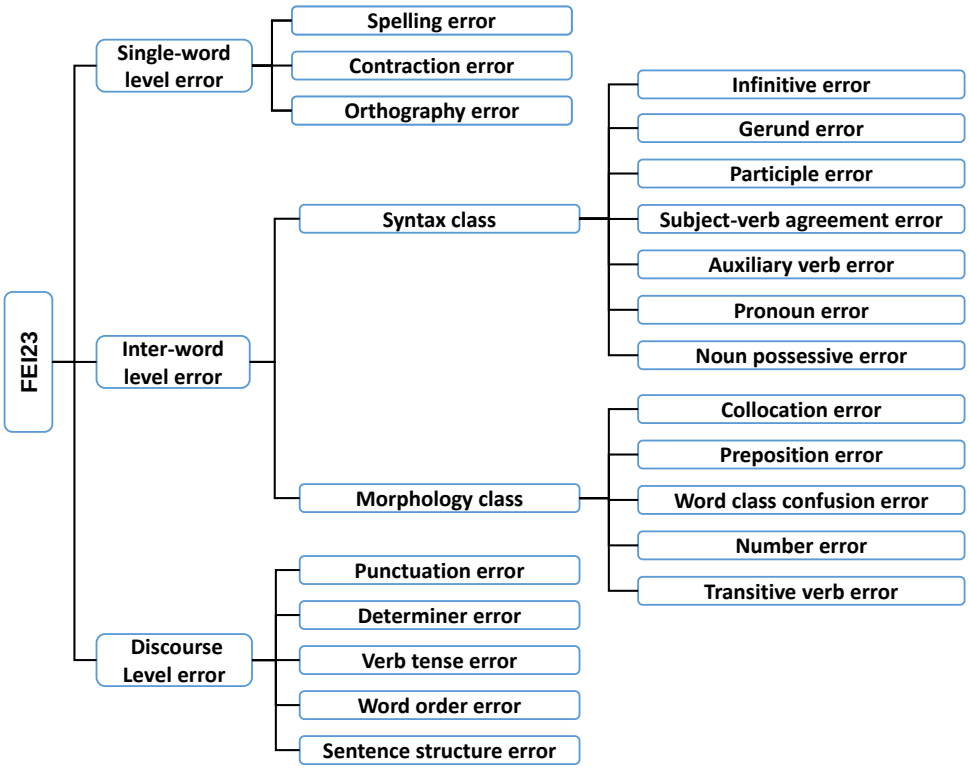

Figure 4: Overview of the FEI23 Error Classification Taxonomy

2.1 Syntax Class
Definition: Errors in sentence grammar and structure.

2.1.1 Infinitive Error
Definition: Errors like missing "to" before a verb in to-infinitives, or unnecessary "to" after modal verbs for zero-infinitives.

Error sentence: I would like going home.
Correct sentence: I would like to go home.

2.1.2 Gerund Error
Definition: Misuse of the verb form that should act as a noun in a sentence.

Error sentence: I enjoy to play.
Correct sentence: I enjoy playing.

2.1.3 Participle Error
Definition: Confusion between participles and ordinary verb tenses.

Error sentence: She has did her homework.
Correct sentence: She has done her homework.

2.1.4 Subject-verb Agreement Error
Definition: The verb does not agree with the subject in number.

Error sentence: They was happy.
Correct sentence: They were happy.

2.1.5 Auxiliary Verb Error
Definition: Misuse of auxiliary verbs such as "do," "have," or modal auxiliaries like "could," "may," "should."

Error sentence: She can sings well.
Correct sentence: She can sing well.

2.1.6 Pronoun Error
Definition: Pronouns do not agree in

number, person, or gender with their antecedents.

> **Error sentence:** Her went to the store.
> **Correct sentence:** She went to the store.

2.1.7 Noun Possessive Error
Definition: Misuse of possessive adjectives and possessive nouns.

> **Error sentence:** This is Johns book.
> **Correct sentence:** This is John's book.

2.2 Morphology Class
Definition: Errors in the structure or form of words, including their grammatical relationships.
2.2.1 Collocation Error
Definition: Atypical word combinations that are grammatically acceptable but not common.

> **Error sentence:** He did a crime.
> **Correct sentence:** He committed a crime.

2.2.2 Preposition Error
Definition: Misuse of prepositional words.

> **Error sentence:** He is good in math.
> **Correct sentence:** He is good at math.

2.2.3 Word Class Confusion Error
Definition: Confusions in part of speech, such as noun/adjective or adjective/adverb confusion.

> **Error sentence:** She is a beauty dancer.
> **Correct sentence:** She is a beautiful dancer.

2.2.4 Number Error
Definition: Confusion in singular or plural form of nouns.

> **Error sentence:** She has many friend.
> **Correct sentence:** She has many friends.

2.2.5 Transitive Verb Error
Definition: Extra preposition after transitive verbs or missing preposition after intransitive verbs.

> **Error sentence:** He gave.
> **Correct sentence:** He gave a gift.

3 Discourse Level Error
Definition: Errors affecting the overall structure, flow, or coherence of a sentence or discourse.
3.1 Punctuation Error
Definition: Errors in the use, omission, or addition of punctuation marks.

> **Error sentence:** She said, "Hello".
> **Correct sentence:** She said, "Hello."

3.2 Determiner Error
Definition: Errors in the use, omission, or addition of determiners.

> **Error sentence:** She bought apple.
> **Correct sentence:** She bought an apple.

3.3 Verb Tense Error
Definition: Incongruities in verb tenses, such as an erroneous tense shift in a compound sentence.

> **Error sentence:** He go to school yesterday.
> **Correct sentence:** He went to school yesterday.

3.4 Word Order Error
Definition: Errors in arranging words in the correct sequence within a sentence.

> **Error sentence:** To school she goes every day.
> **Correct sentence:** She goes to school every day.

3.5 Sentence Structure Error
Definition: Errors affecting the overall grammatical structure of a sentence.

> **Error sentence:** When he comes, I will leave before he arrives.
> **Correct sentence:** When he comes, I will leave.

H Taxonomy-Specific Issues Affecting Rationality Metrics

This appendix provides a detailed analysis of taxonomy-specific issues that impact the rationality metrics assessed in our study.

POL73 contains categories such as Indefinite article incorrect and Determiners, which often capture the same errors, leading to overlapping classifications. Similarly, Inappropriate words but correct word class and General semantic confusion lack clear differentiation, reducing mutual exclusivity. TUC74 introduces ambiguous categories, such as Simple predicate missing: be vs. BE missing, and Surrogate subject missing: there/it vs. Omission of surrogate subject, creating annotation inconsistencies. FEI23, by structuring categories based on cognitive levels, ensures high exclusivity by clearly defining error boundaries and reducing ambiguity.

POL73 includes categories like New creations but lacks fundamental categories such as Punctuation error and Orthography error, slightly reducing its coverage. TUC74 performs worse in coverage as it omits not only Punctuation error and Orthography error but also Spelling error. This forces annotators to rely excessively on the Other category, significantly lowering its coverage score. BRY17, by defining error types based on both part-of-speech and token-level operations, ensures a broader representation of errors, leading to the highest coverage among all taxonomies.

TUC74 demonstrates extreme specificity by defining five separate categories for errors related to temporal conjunctions, such as Misplacement of While. This level of granularity results in sparsely populated error categories, skewing distributional balance. POL73 takes the opposite approach, overgeneralizing error types. The category Inappropriate words and incorrect word class encompasses a wide range of unrelated grammatical issues, resulting in an overly broad classification that reduces distinction between different error types. FEI23 adopts a hierarchical classification strategy at three levels—Single-word level, Inter-word level, and Discourse level—ensuring balanced categorization without excessive fragmentation or aggregation.

POL73 and TUC74 include overly technical or abstract error categories, such as Phonetic similarity within English and Misuse of end-of-the-road predicates, making them difficult to interpret and apply consistently. BRY17 and FEI23, by contrast, use straightforward category labels with simple lexical definitions, improving both human annotation efficiency and model interpretability. Their classification schemes facilitate direct mapping to computational models, reducing the complexity of automated grammar error detection.

These taxonomy-specific issues highlight the challenges in designing effective error classification taxonomies and their impact on the rationality metrics assessed in our study.

I Ablation Study on Category Fusion

To examine the impact of classification granularity on four metrics, we conduct an ablation study by merging specific error categories. The specific taxonomy structures before and after fusion are detailed in the Appendix J. To ensure control for variability in model predictions, all results in this study are only derived from ChatGPT-4o. The experimental results are shown in Table 5.

Our findings indicate that the fusion of error categories leads to a consistent increase in Coverage across all classification taxonomies. This is expected, as the newly formed parent category inherently encompasses a broader scope than its individual subcategories. By expanding the definition of error categories, previously unclassified errors are now incorporated into the taxonomy, leading to a higher coverage score.

Conversely, Balance decreases across all taxonomies after category fusion. This decline can be attributed to the inherent structure of the original taxonomies: error categories with broad definitions already contained a substantial number of samples prior to fusion. Merging smaller categories into these broader ones exacerbates the data imbalance, further skewing the error distribution. As a result, the overall balance of the taxonomy deteriorates.

The impact on Exclusivity varies depending on the specific taxonomy and the relationships between merged categories. For instance, in BRY17, the pre-fusion taxonomy contained overlapping error categories such as *Particle* and *Preposition*. By merging these, the overlap is eliminated, leading to an increase in exclusivity. However, in FEI23, pre-fusion error types such as *Contraction error*, *Spelling error*, and *Orthography error* were mutually exclusive and did not overlap with other categories. After merging them into the broader *Single-word level error*, overlaps emerged with other categories like *Pronoun Error*, resulting in a decrease in exclusivity.

Taxonomies	Exclusivity ($\tau = 0.7$) ↑	Coverage ↑	Balance ↑	Usability (Macro F1 / Micro F1) ↑
POL73	0.842	0.698	**0.687**	0.301 / 0.478
POL73 (Fusion)	**0.886**	**0.715**	0.662	**0.371 / 0.524**
TUC74	0.703	0.160	**0.210**	0.061 / 0.099
TUC74 (Fusion)	**0.774**	**0.183**	0.178	**0.131 / 0.113**
BRY17	0.921	0.979	**0.829**	0.610 / 0.760
BRY17 (Fusion)	**0.932**	**0.984**	0.786	**0.618 / 0.800**
FEI23	**0.877**	0.924	**0.878**	**0.631 / 0.743**
FEI23 (Fusion)	0.830	**0.925**	0.834	0.616 / 0.719

Table 5: Ablation Study on the Impact of Merging Error Categories (ChatGPT-4o). Taxonomies labeled with Fusion indicate error classification taxonomies after category merging.

Similarly, Usability is influenced by classification granularity, as broader error categories reduce the model's ability to distinguish fine-grained errors. Merging categories weakens decision boundaries, increasing intra-category variability and making classification more ambiguous. While higher Exclusivity can enhance usability by reducing overlaps, it is not the sole factor—when fusion introduces ambiguity, usability declines due to the model's reduced classification accuracy. Our results confirm that category fusion affects usability by altering decision boundaries and classification effectiveness.

Our ablation study on category fusion demonstrates that increasing classification granularity leads to higher Coverage by broadening error definitions, but decreases Balance due to exacerbated data imbalance. The effect on Exclusivity depends on pre-fusion category overlap, improving when redundant categories are merged but declining when new overlaps emerge. The impact on Usability varies depending on how category fusion reshapes classification boundaries—some fusions improve clarity, while others introduce ambiguity that weakens classification effectiveness. Our findings highlight an important challenge in error taxonomy design: both excessive simplification and over-fragmentation can negatively impact classification taxonomies. Constructing an effective taxonomy is highly complex and cannot be achieved solely through empirical intuition, underscoring the importance of rigorous evaluation metrics in assessing classification taxonomies.

J Error Category Fusion Details

J.1 POL73

Before Fusion:

2.2 Verb Phrase
Definition: Errors related to verb use, including tense, form, and omission.
2.2.1 Omission of Verb
Definition: Errors involving the omission of a necessary verb, including main verbs and auxiliary verbs such as forms of "be".

> **Error sentence:** He to school yesterday.
> **Correct sentence:** He went to school yesterday.

2.2.2 Use of Progressive Tense
Definition: Errors in forming or using the progressive tense, including omission of the auxiliary verb "be", incorrect use of the "-ing" form, or substitution of the progressive where it is not expected.

> **Error sentence:** He is play football.
> **Correct sentence:** He is playing football.

2.2.3 Agreement of Subject and Verb
Definition: Errors where the subject and verb fail to agree in number, person, or tense.

> **Error sentence:** I didn't know what it is.
> **Correct sentence:** I didn't know what it was.

After Fusion:

2.2 Verb Phrase
Definition: Errors related to verb use, including tense, form, and omission.

> **Error sentence:** He to school yesterday.
> **Correct sentence:** He went to school yesterday.

J.2 TUC74

Before Fusion:

1.1 Missing Parts
Definition: Errors involving the omission of essential sentence components necessary for grammatical completeness and structural clarity.
1.1.1 Surrogate Subject Missing: there and it
Definition: Missing placeholder subjects "there" or "it," which are required in English sentence structure.

> **Error sentence:** Was a riot last night.
> **Correct sentence:** There was a riot last night.

1.1.2 Simple Predicate Missing: *be*
Definition: Missing the simple predicate "be," which is required when the predicate consists of adjectives or noun phrases.

> **Error sentence:** My sisters very pretty.
> **Correct sentence:** My sisters are very pretty.

1.1.3 Object Pronoun Missing
Definition: Missing object pronouns (direct or indirect) in places where verbs require them.

> **Error sentence:** Donald is mean so no one likes.
> **Correct sentence:** Donald is mean so no one likes him.

1.1.4 Subject Pronoun Missing
Definition: Missing subject pronouns where they are required, especially after subordinate conjunctions.

> **Error sentence:** He worked until fell over.
> **Correct sentence:** He worked until he fell over.

After Fusion:

1.1 Missing Parts
Definition: Errors involving the omission of essential sentence components necessary for grammatical completeness and structural clarity.

> **Error sentence:** My sisters very pretty.
> **Correct sentence:** My sisters are very pretty.

J.3 BRY17

Before Fusion:

2.1.3 Conjunction (CONJ) [R/U/M]
Definition: Errors in the use, omission, or addition of conjunctions.

> **Error sentence:** She likes apples oranges.
> **Correct sentence:** She likes apples and oranges.

2.1.6 Particle (PART) [R/U/M]
Definition: Errors involving particles, such as their incorrect placement, omission, or addition.

> **Error sentence:** He got up into the car.
> **Correct sentence:** He got into the car.

2.1.10 Preposition (PREP) [R/U/M]
Definition: Errors in the use, omission, or addition of prepositions.

> **Error sentence:** He is good math.

Correct sentence: He is good at math.

After Fusion:

2.1.8 Function Word (FUNC:WORD) [R/U/M]
Definition: Errors in the use, omission, or addition of function words. These errors affect sentence structure and meaning by altering how words and phrases connect.

Error sentence: She likes apples oranges.
Correct sentence: She likes apples and oranges.

J.4 FEI23

Before Fusion:

1 Single-word Level Error
Definition: Errors confined to a single word, typically involving spelling, contractions, or capitalization.
1.1 Spelling Error
Definition: Incorrect spelling of a word.

Error sentence: She recieved the letter.
Correct sentence: She received the letter.

1.2 Contraction Error
Definition: Incorrect use or formation of contractions.

Error sentence: I isn't ready.
Correct sentence: I'm not ready.

1.3 Orthography Error
Definition: Errors in capitalization or other writing conventions.

Error sentence: he went to school.
Correct sentence: He went to school.

After Fusion:

1 Single-word Level Error
Definition: Errors confined to a single word, typically involving spelling, contractions, or capitalization.

Error sentence: She recieved the letter.
Correct sentence: She received the letter.