When One Label Is Not Enough: Rethinking Error Classification in Grammatical Error Detection

Anonymous ACL submission

Abstract

For several years, Grammatical Error Detection (GED) has gone beyond binary classification to categorize errors according to a predefined typology. Whereas an erroneous token could potentially be described with several error labels, most existing approaches treat error classification as a multiclass problem, assigning a single label to each erroneous token. In this paper, we critically examine classification methods in the GED literature and argue that annotation schemes should explicitly model the multilabel nature of errors rather than force a single-label assignment. This survey reveals that most studies overlook the multilabel nature of the problem, while others address it indirectly through alternative strategies such as mutually exclusive typologies, label hierarchies, residual classes, and hybrid systems. Conversely, explicit multilabel modelling remains relatively rare and presents mixed results. We discuss several factors that contribute to this situation, including the scarcity of multilabel-annotated learner corpora, the cost and complexity of annotation, and the methodological challenges posed by label dependency and label imbalance. Finally, we outline directions for future research, including the systematic analysis of multilabel properties in learner corpora and the exploration of multilabel learning methods developed in other areas of machine learning. Our findings highlight the importance of aligning modeling choices with the structural properties of learner errors in order to improve the interpretability and analytical usefulness of GED systems.

1 Introduction

Grammatical Error Detection (GED) is a well-established research area in NLP that aims to automatically identify errors — grammatical or otherwise — present in a text (Yuan et al., 2021; Bryant et al., 2023). Since Rei and Yannakoudakis (2016), GED has generally been defined as a sequence labelling task where erroneous tokens are distinguished from correct ones. However, such binary information is insufficient for detailed error analysis (Yuan et al., 2021). For this reason, research in GED has progressively moved beyond binary classification by expanding the set of labels that can be assigned to erroneous tokens through predefined error typologies.

With this shift towards multi-category classification, GED has attracted particular interest in second language acquisition research, where the categorization of errors enables the providing of *corrective feedback* to learners (Casademont Moner and Volodina, 2022). However, it is important to note that this extended classification can be parameterized in two different ways: as multiclass classification (MCC) (also referred to as *single-label*), in which each element can receive only one label (see Example 1a), or as multilabel classification (MLC), in which each element may be assigned several error labels simultaneously (see Example 1b) (García-Pedrajas et al., 2024). In other words, categories are mutually exclusive in MCC, whereas they can be mutually inclusive in MLC.

As an example of this distinction, hereafter is the processing of the erroneous string "Problem" in the sentence "Cars and buses are dangerous and cause **Problem* on the street"[1], which involves two error types:

(1a) Multiclass classification: choice between a capitalization error and a number error;

(1b) Multilabel classification: capitalization AND number error.

In practice, the choice between these parameterizations is often driven by implementation decisions and, for supervised approaches, by the annotation scheme adopted in the training corpus. Nevertheless, at least for error classification, using MCC

[1]This is an authentic sentence from the W&I+LOCNESS corpus (Bryant et al., 2019).

raises several concerns. Returning to Example 1, we have seen that two different error types objectively affect "*Problem" simultaneously. However, MCC requires assigning a single label, which in this case means that one of the two error types is overlooked. Yet, given that uncorrected errors are likely to fossilize in learner grammars (Shahid et al., 2023), partial error categorization in educational applications of GED may pose a similar risk. By failing to account for the full set of error types affecting a token, such systems may negatively affect the quality of corrective feedback and lead learners to incorrectly assume that unflagged aspects of the token are error-free. This issue is all the more critical in inflectional and agglutinative languages, where morphological richness makes multiple errors within a single token more likely (Kempfert and Köhn, 2018; Korre et al., 2021; Katinskaia et al., 2022; Uz and Eryiğit, 2023).

Moreover, the partial labeling induced by an approach misaligned with the nature of the data may also affect model performance. Although this issue has rarely been discussed in the GED literature, several studies in image classification have shown that annotating intrinsically multilabel data under a multiclass scheme degrades classification performance (Xu et al., 2019; Yun et al., 2021). This can be explained by the fact that when instances may legitimately belong to several categories, models tend to produce multilabel predictions (Yun et al., 2021). Yet, multiclass-annotated corpora unjustifiably penalize such predictions, thereby introducing structural bias into the learning process (Verelst et al., 2022; Zhou et al., 2022). Based on these observations, Yun et al. (2021) even argue that the mismatch between the annotation scheme and the structure of the data may itself constitute a form of annotation error.

Despite the potentially negative consequences of this mismatch, our comprehensive survey of the classification methods in GED literature reveals that the majority of authors adopt a multiclass approach that does not reflect the inherently multilabel nature of the error classification task. The aim of this paper is therefore to investigate why this paradigm choice is made and how researchers apply a MCC framework to data in which a single instance may legitimately belong to multiple categories. In doing so, we seek to clarify methodological choices in GED and point out potential directions for more accurate error modeling.

Contributions. This position paper makes four main contributions: (1) we identify a methodological mismatch between the nature of the data and the classification paradigms commonly used in GED, based on a comprehensive survey of the field; (2) we propose a typology of strategies used in the GED literature to handle error classification; (3) we provide a comprehensive and critical survey of the state of the art in models developed to perform tasks of classifying grammatical errors made by foreign language learners; (4) we discuss multilabel methods that could be applied to GED.

2 Survey of State-of-the-Art Approaches

To investigate how GED models are parameterized, this section provides, to the best of our knowledge, a comprehensive survey of the state of the art in automated categorization of grammatical errors[2] produced by foreign language learners. In order to determine whether the models presented in these studies are configured for MCC or MLC, we examined the use of mutually exclusive versus non-exclusive classes and, in the latter case, the implementation and conceptual choices made by the authors. Based on this analysis, we identify five main approaches: (1) MCC based on mutually exclusive typologies, (2) MCC without multilabel awareness, (3) MCC despite multilabel awareness, (4) partially MLC, and (5) fully MLC. Table 1 illustrates the performance of these approaches through an analysis of the erroneous word "*dictronary" (expected form: "dictionaries"), as it appears in the sentence "In comparison, nowadays electronic **dictronary* are much more multi-functional" [3].

It should be noted that classifying the reviewed articles into these categories was not always straightforward. As we observed that terminological confusion sometimes exists between the terms multilabel and multiclass[4], we considered that the mere use of these terms was insufficient evidence to determine implementation choices with certainty. Then, in the absence of explicit textual evidence (i.e., a clear statement of the parameterization), we also considered, when relevant, the lineage of the model and the type of classification

[2]We decide to exclude studies in Grammatical Error Correction (GEC) that employ GED as an intermediate step towards correction and in which error classification is not evaluated (e.g. Omelianchuk et al. (2020)).

[3]Example from NUCLE corpus (Dahlmeier et al. (2013).

[4]For instance, Ehnroth and Park (2023) state: "We define the multilabel classification of tokens as the task of classifying each token in a sentence into one of the five types of error [in our typology].".

Approach	Definition	Expected treatment	Actual treatment	Management of the multilabel aspect
MCC based on mutually exclusive typology	Classes are mutually exclusive by design; MLC is impossible	Spelling and Number error	Noun error	Not applicable
MCC without multilabel awarness	Single-label output with no acknowledgment that MLC could apply		Spelling error	Non-existent without indications
MCC despite multilabel awarness	Single-label output despite acknowledging that MLC could apply		Spelling error	Non-existent with indications
Partially MLC	MLC is possible for certain error types (typically morphological), but not others		Spelling error	Non-existent with a warning for single-label cases OR Existent
Fully MLC	Multiple labels can be assigned to any error		Spelling and Number error	Existent

Table 1: The five approaches to GED categorization and their treatment of **dictronary* — a token bearing both a spelling and a number error — in the sentence "In comparison, nowadays electronic {*dictronary* are much more multi-functional." extracted from the NUCLE corpus (Dahlmeier et al., 2013).

used in the training corpus. Furthermore, whenever possible, we empirically tested the models and their outputs. Table 3, in the appendices, summarizes the characteristics of the articles discussed in this section according to these criteria.

2.1 Multiclass Classification Based on Mutually Exclusive Typologies

Among all studies scrutinized, only two rely on typologies that are, by design, strictly mutually exclusive as they rely solely on grammatical classes. Fioravanti et al. (2024) construct a typology consisting of eight error types obtained by combining a grammatical class with a morphosyntactic error (e.g., incorrect token selection, *etc.*). As regards the classes used in the SErCl model (Choshen et al., 2020), they are defined as combinations of the erroneous grammatical class and the expected grammatical class.

It is important to note that, in order to construct such typologies, these authors deliberately focus on specific types of errors (morphosyntactic errors in the case of Fioravanti et al. (2024) and syntactic errors in Choshen et al. (2020)). As a consequence, these models can only provide partial information about the nature of the errors and fail to capture some error types. For instance, the model of Fioravanti et al. (2024) will assign a single label indicating an incorrect ending when gender and number errors co-occur. Similarly, if an error is unrelated to grammatical class, the model of Choshen et al. (2020) will not identify it.

2.2 Multiclass Classification Without Multilabel Awareness

All the remaining studies rely on typologies whose labels are, in principle, combinable, yet most adopt implementation choices that do not account for the intrinsic multilabel nature of the data. This is explicitly the case in Mizumoto et al. (2024), who use a generative model for GED and clearly specify in their prompt that each erroneous token must receive exactly one label.

In many other studies, however, this parameterization is only implicit. This is notably the case for ERRANT (Bryant et al., 2017), one of the most widely used symbolic models for GED, which does not address the issue of MLC despite relying on non-exclusive error categories.[5] To determine how the model is parameterized, we evaluated ERRANT on three erroneous variants of the word "grammatical" in the sentence "This is a grammatical sentence." (Table 2). While capitalization and spelling errors are correctly detected individually, their co-occurrence within the same token leads ERRANT to annotate only the spelling error, suggesting that the model effectively operates under an MCC scheme.

Token	Expected	ERRANT
Grammatical	R:ORTH	R:ORTH
gra**mam**tical	R:SPELL	R:SPELL
Gra**mam**tical	**R:ORTH**	R:SPELL
	R:SPELL	

Table 2: Expected and actual classification produced by the ERRANT model of the three erroneous writings of the "grammatical" token.

A similar implicit parameterization can be inferred for the symbolic approaches of Náplava et al.

[5]Each error in this typology type combines information about the operation required to normalize the error (e.g., missing token, *etc.*) with information about the grammatical category or the nature (e.g., spelling, *etc.*) of the error.

(2022), Katinskaia et al. (2022), Uz and Eryiğit (2023), and Choshen et al. (2021), as well as the discriminative model of Yuan et al. (2021), all of which rely on typologies derived from ERRANT. Although these studies do not explicitly discuss multilabel classification, their lineage from ERRANT strongly suggests an MCC parameterization.

For their part, Kara et al. (2023) and Nguyen et al. (2024) train their discriminative models on corpora in which each token is annotated with a single label. Consequently, their models can only perform MCC. Finally, the unique indication regarding the parameterization use by Bhattacharyya and Bhattacharya (2025) is their use of the term "multiclass" — rather than "multilabel" — to describe the token-level classification task.

Overall, these studies either explicitly adopt MCC or strongly suggest such a parameterization without discussing its implications. Yet, this absence of explicit consideration is particularly problematic in educational applications, where users of the models may incorrectly assume that the predicted label fully characterizes the error affecting a token.

2.3 Multiclass Classification Despite Multilabel Awareness

Unlike the studies discussed above, the works examined in this subsection explicitly acknowledge that a single token may involve several error categories, while nevertheless constraining classification to MCC. The distinction is important: although both configurations ultimately produce the same multiclass behavior at inference time, these studies make the model's assumptions and limitations explicit. As a result, users are less likely to interpret the predicted label as an exhaustive characterization of the error affecting a token.

In their symbolic models, Korre et al. (2021), Yoon et al. (2023), and Kosakin et al. (2024) explicitly acknowledge that a single token may involve several error types, while nevertheless constraining classification to MCC. Whereas Korre et al. (2021) do not specify how such multilabel cases are resolved, Yoon et al. (2023) and Kosakin et al. (2024) rely on label priority schemes to enforce a single-label prediction. The authors also differ in how they justify this choice. Whilst Korre et al. (2021) presents it as a limitation to be addressed in future work, Yoon et al. (2023) describes it as a deliberate simplification intended to facilitate the evaluation and analysis of error distribution. For their part, and although Kosakin et al. (2024) introduce a multilabel-annotated corpus, they provide no explicit justification for adopting MCC in their automatic classification model.

Another way of acknowledging the multilabel nature of data without implementing MLC it is to group all multilabel cases into a residual category. For instance, Oshin et al. (2023) assign indifferently all multilabel instances in their training corpus to a category labeled *multiple errors*. While this strategy has the advantage of recognizing that some tokens may involve several error types, it considerably limits the analyses that can be conducted on the affected tokens and may perturbate the model with a noisy category. Its usefulness therefore depends on factors such as the size of the typology and the prevalence of multilabel instances in the data. In the case of Oshin et al. (2023), the resulting information loss remains relatively limited, as their typology contains only three additional categories.

2.4 Partially Multilabel Classification

Some models allow MLC only in specific cases, while the majority of predictions remain multiclass. The symbolic model proposed by Rozovskaya (2022), for example, operates in two stages by establishing a priority order among the labels in its typology. First, the model checks whether the token contains a non-morphological error (e.g., spelling, punctuation, *etc.*) or whether the error concerns a preposition, an adverb, or a pronoun. As soon as an error of this type is detected, it is annotated and the analysis stops. Otherwise, if none of these error types apply, the model proceeds to analyze morphological errors affecting nouns, verbs, and adjectives, in which case it assigns as many labels as necessary. Similarly, Kempfert and Köhn (2018) define a hierarchy among the labels in their typology. MLC is applied only when the symbolic model detects an error related to grammatical class, in which case several relevant subtypes may be assigned within that broader category.

Alongside these two models, Nasyrova and Sorokin (2025) and Qin et al. (2023) constitute borderline cases within our typology, as they implement only highly restricted forms of partial MLC. Specifically, Nasyrova and Sorokin (2025) allow multiple *insertion* labels between tokens, while cases involving several labels on the same token are removed during preprocessing. Their discrim-

inative model is therefore effectively trained and applied under an MCC scheme at the token level. Similarly, Qin et al. (2023) use two separate symbolic classifiers, one for grammatical errors and the other for phonetic errors. However, based on the model's lineage from ERRANT and the excerpt of the grammatical decision tree presented in their article, classification appears to remain multiclass within each individual classifier. Consequently, except for these highly specific configurations, both approaches ultimately fall within the category of MCC without multilabel awareness.

2.5 Fully Multilabel Classification

In contrast to the numerous studies discussed so far, the models presented in this section explicitly acknowledge the multilabel nature of their data and implement classification strategies accordingly. Reviewing these, we outline the two main MLC approaches that have been applied to GED.

2.5.1 Algorithm adaptation

One way to perform MLC consists in adapting existing binary or multiclass classification methods to this specific task (Herrera et al., 2016). The three studies presented in this subsection can be viewed as instances of this approach, in that they both adapt models originally designed for MCC to multilabel settings with a few number of modifications.

Belkebir and Habash (2021), who propose another symbolic model derived from ERRANT, attempt to extend the rule-based approach used in MCC by ERRANT and others to a multilabel setting. To this end, they design a set of rules for classifying tokens into the 26 types of errors defined in their typology. The main difference between this model and those discussed previously lies in the fact that the system does not stop once an error has been detected. Thus, theoretically, this approach allows all possible combinations of labels to be assigned. In practice, however, the authors report that fully combinatorial MLC could not be achieved due to efficiency constraints and the excessive growth in dimensionality.

Following a similar principle, Noreskal et al. (2023) and, building on their work, Yan et al. (2025) apply discriminative models based on an architecture similar to that used in MCC. The main modification introduced to enable multilabel prediction is that, instead of producing a single output corresponding to one class, each classification head of the neural network is specialized in predicting a specific label. However, using this approach, Noreskal et al. (2023) report accuracy scores ranging from 54.5% at best to as low as 4.5%. One explanation proposed by the authors for these modest results is the imbalance of labels in the training corpus. This hypothesis is supported by Yan et al. (2025), who show that cleaning the dataset and artificially augmenting underrepresented categories significantly improves the prediction performance for these labels. It should be noted that label imbalance is a well-known issue in classification tasks; we discuss it in greater detail in Section 3.3.

2.5.2 Data transformation

Rather than adapting algorithms to the data, another possible strategy is to transform the data so that they can be processed by algorithms originally designed for binary or multiclass classification (Herrera et al., 2016; Bogatinovski et al., 2021).

The *binary relevance* method — the most widely used technique in MLC (Herrera et al., 2016) — transforms a multilabel problem into several independent binary classification tasks. For each label, the presence of the label is treated as a positive instance and its absence as a negative instance, independently of the other labels. As a result, the label representation in the training corpus is modified for each classifier, and a separate binary classifier is trained for each label, leading to as many classifiers as there are labels in the typology (Herrera et al., 2016; Bogatinovski et al., 2021). Both Huang and Wang (2016) and Wang et al. (2023)[6] apply the *binary relevance* approach to GED by training a separate model for each error type (Bi-LSTMs in the case of Huang and Wang (2016) and discriminative Transformers for Wang et al. (2023)). The outputs of these different models are then combined into a final prediction vector. One of the main advantages of this approach is that it helps mitigate the problem of high output dimensionality (Huang and Wang, 2016). However, it may also exacerbate label imbalance (see Section 3.3), thereby reducing the prediction quality for less frequent labels (Herrera et al., 2016; Bogatinovski et al., 2021). Huang and Wang (2016) encounter precisely this issue with one of their cat-

[6] It should be noted that, although effectively performing MLC, Wang et al. (2023) appear to conceptualize the task as multiclass.

egories but, instead of applying commonly used rebalancing or resampling strategies, they choose to merge the problematic error type with one of the three remaining categories. Because the two merged categories differ in the number of tokens involved, they can later be distinguished during post-processing. Nevertheless, the evaluation results reported by both groups of authors indicate that this approach still requires further improvement. Huang and Wang (2016) report, at best, an accuracy of 0.46, a precision of 0.34, a recall of 0.37, and an F1 of 0.32. For Wang et al. (2023), the model achieves a precision of 35.34, a recall of 68.51, and an F1-score of 46.63 for coarse-grained categories, but performance drops considerably for fine-grained typologies, with scores of 21.16, 36.45, and 26.77 respectively.

Another transformation-based approach is the *label powerset* method, which treats MLC as a multiclass problem by collapsing the combinatorial space of possible label combinations into a broad set of categories (Herrera et al., 2016; Bogatinovski et al., 2021). This can be done either *a priori* by generating classes for all possible label combinations defined by the typology (Herrera et al., 2016), or more pragmatically by creating classes only for the combinations that appear in a multilabel-annotated training corpus (Bogatinovski et al., 2021). The main drawback of the first strategy is the combinatorial explosion that arises when the typology contains more than a few labels[7]. While the second approach avoids this issue, it limits the model's predictive capacity, as the classifier cannot assign label combinations that do not appear in the training data (Bogatinovski et al., 2021). Moreover, both implementations may reduce the number of examples available for each category. In the case of the a priori approach, some classes may even be entirely unrepresented in the training corpus. This will negatively affect the model's generalization ability (Herrera et al., 2016) and may lead to underfitting (Bogatinovski et al., 2021). Alhafni et al. (2023) adopt the pragmatic version of the *label powerset* approach for GED and obtain a total of 43 classes, including classes corresponding to combinations of labels. Given that the original typology contains 32 error tags, it is possible to calculate that, in theory, this typology allows 2^{32} possible combinations (i.e., 4,294,967,296 distinct classes). Although the majority combinations within these theoretical 2^{32} classes are undoubtedly not possible, it is likely that the 43 classes used in practice would not cover all possible multilabel cases in new instances. In terms of performance, the model achieves, at best, a precision of 0.57, a recall of 0.48, and an $F_{0.5}$-score of 0.53.

[7]"Theoretically, 2^k different label combinations can exist into an MLD with k labels." (Herrera et al., 2016).

3 Issues, Challenges, and Potential Solutions for Multilabel Classification

To better understand why MLC remains underused compared to MCC in GED, this section revisits three of the main issues that influence the application and development of MLC methods (Bogatinovski et al., 2021): training corpora, label dependency and label imbalance. While the label dependency is specific to MLC, the two others are well known and widely studied in binary and multiclass classification tasks. However, as we will show, these challenges present particular characteristics in multilabel settings and require adaptations of the solutions originally developed for binary and multiclass tasks (Herrera et al., 2016; Bogatinovski et al., 2021).

3.1 Corpora and annotation cost

A first factor that may explain the limited use of MLC in GED is the near absence of learner corpora annotated using this paradigm that could serve as training data for models. Among the most commonly used corpora for English as a foreign language, only the National University of Singapore Corpus of Learner English (NUCLE) implements MLC (Dahlmeier et al., 2013). However, it is worth noting that, for the sake of standardization, this corpus was automatically reannotated using ERRANT — which applies MCC (see Section 2.2) — during the Building Educational Applications (BEA) shared task (Bryant et al., 2019). The Lang-8 Corpus (Mizumoto et al., 2011) and the Cambridge English Write & Improve + LOCNESS (W&I) corpus (Bryant et al., 2019) were also automatically annotated using ERRANT on this occasion (Bryant et al., 2019). As for the Cambridge Learner Corpus — and the First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011), which derives from it — the article introducing the corpus explicitly states that the use of multiple labels for a single error was deliberately avoided (Nicholls, 2003).

This lack of suitable corpora consequently hin-

ders the development of models that require task-specific training. However, constructing such corpora is not trivial. While corpus annotation for multiclass tasks is already known to be time-consuming, the process becomes even more complex in a multilabel setting (Xu et al., 2022; Su et al., 2026; Arroyo et al., 2023). The reason is that, as the number of classes increases, annotators face increasing difficulty in identifying all categories applicable to a given token (Cole et al., 2021). Moreover, MLC reduces the likelihood of agreement between annotators and complicates the computation of inter-annotator agreement (Marchal et al., 2022).

This difficulty in fully annotating corpora is well known in other fields, such as document retrieval and image classification, where positive-unlabeled (PU) learning has been proposed as a possible solution. In this approach, labels are either annotated as positive or treated as unlabeled, meaning that unannotated labels may correspond either to missed positives or to true negatives (Chen et al., 2026). Although this approach reduces annotation costs, recent work shows that its effectiveness strongly depends on data type, label frequency, and labeling mechanism, and that it performs poorly on highly sparse data (Chen et al., 2026). Since most authentic learner corpora — including those used in GED — are inherently sparse (Bogatinovski et al., 2021; Chen et al., 2026), fully multilabel-annotated corpora currently remain one of the most reliable solutions.

3.2 Label dependency

Label dependency refers to the idea that certain labels may be related to one another (Bogatinovski et al., 2021). In other words, the presence of some labels increases the probability of the presence of another correlated label. Such dependencies may be unconditional, meaning that labels are intrinsically correlated (e.g., some subject-verb agreement errors often lead to a number error), or conditional, when labels are only related within a specific context (e.g., in a given context, a token simultaneously involves both a gender and a number error) (Herrera et al., 2016).

The existence of correlations between labels is not problematic in itself; rather, it is an inherent characteristic of MLC. The difficulty arises because some methods used for MLC (e.g., *binary relevance*) rely on an independence assumption and, therefore, do not account for the dependencies between labels (Huang et al., 2024). Unless extreme cases where labels are truly independent, this assumption can negatively affect model effectiveness and robustness (Bogatinovski et al., 2021; Cheng et al., 2023; Huang et al., 2024). By contrast, in algorithmic adaptation approaches (e.g., the models proposed by Noreskal et al. (2023) or Yan et al. (2025)), classifiers are modified to directly model such dependencies (Bogatinovski et al., 2021). Similarly, by creating a label for each possible set of error types, the *label powerset* method implicitly captures these relationships (Herrera et al., 2016).

A first step toward determining which method best fit one's data is to measure the extent to which the corpus contains instances likely to receive multiple labels. Two commonly used metrics for this purpose are label cardinality and label density. Label cardinality measures the average number of labels assigned to each instance in the corpus. Thus, if the cardinality equals 1, this indicates that all instances are associated with a single label and that MCC is sufficient. Conversely, the higher the value rises above 1, the more instances in the corpus tend to accumulate multiple labels. Label density corresponds to a normalized version of label cardinality, ensuring that the resulting measure is not influenced by the size of the typology (Herrera et al., 2016).

3.3 Label imbalance

The existence of correlations between labels can also exacerbate the well-known classification issue of label imbalance. This phenomenon arises when, within the training corpus, some labels are highly represented while others occur only rarely (Herrera et al., 2016)[8]. As a consequence, labels that are underrepresented during training are more likely to be misclassified than those that are more frequent (Herrera et al., 2016; Noreskal et al., 2023; Yan et al., 2025).

In binary and multiclass classification, researchers commonly address this imbalance through resampling methods. These approaches either modify the training corpus by removing tokens belonging to dominant classes, or add tokens labeled with minority classes — by cloning existing instances or artificially synthesizing new

[8] Note that label imbalance is intrinsic to GED, as correct tokens necessarily outnumber erroneous ones. Yet, among the different error types, imbalance may also arise when certain labels are underrepresented in the corpus.

data[9] — or combine both strategies (Herrera et al., 2016). However, in MLC, it may occur that overrepresented labels are correlated with underrepresented ones. This phenomenon, known as *concurrence among imbalanced labels*, significantly reduces the effectiveness of these methods (Herrera et al., 2016). Consequently, when correlations between labels are observed, it may be preferable to rely on classifier adaptation methods — which modify the classifier architecture so that it explicitly accounts for the imbalanced nature of the data (Herrera et al., 2016) — rather than on resampling strategies.

Independently of label dependency, MLC tends to be more affected by label imbalance because the number of underrepresented categories is typically larger (Piras et al., 2021). Furthermore, MLC methods may themselves exacerbate this imbalance. Given that this imbalance problem becomes more significant as the number of classes increases, approaches such as *label powerset* can rapidly become problematic (Herrera et al., 2016; Bogatinovski et al., 2021). Similarly, the "one-vs-all" formulation underlying the *binary relevance* method modifies the original distribution of labels by further diluting the proportion of positive instances, as it increases the number of instances considered negative (Herrera et al., 2016).

A first good practice for addressing this challenge is to explicitly measure the extent of imbalance in the training corpus. Several metrics exist for this purpose. For instance, the *IRLbl* metric measures the degree of imbalance for a specific label by comparing its frequency with that of the most frequent label in the corpus (Herrera et al., 2016). Two additional measures derived from *IRLbl* can also be computed: *MaxIR*, which captures the maximum imbalance by calculating the ratio between the most frequent and the least frequent label, and *MeanIR*, which corresponds to the average of the *IRLbl* values across all labels (Herrera et al., 2016). Finally, the *SCUMBLE* metric measures the extent to which the phenomenon of *imbalanced label concurrence* is present in the corpus (Herrera et al., 2016).

4 Conclusion

In this paper, we examined how grammatical error classification is approached in the literature on Grammatical Error Detection. While learner errors can often involve several linguistic dimensions simultaneously, our survey shows that most existing approaches rely in practice on multiclass classification. A range of strategies has been proposed to reconcile this paradigm with potentially multilabel data, including mutually exclusive typologies, label hierarchies, residual classes, or partially multilabel systems. Although these pragmatic solutions simplify modeling, they lack theoretical foundations and data adherence, as they can still assign a single label to a multi-dimensional error. In some contexts, such as identifying a dominant error type, this simplification may be sufficient. However, in settings where detailed linguistic analysis (e.g. establishing error profiles to learners) or reliable pedagogical feedback is required, it may hinder the ability of GED systems to fully represent the complexity of learner errors.

Our analysis also highlights several reasons why MLC remains relatively uncommon in GED. In particular, the scarcity of multilabel-annotated learner corpora and the high cost and complexity of collecting new annotations constitute important barriers. In addition, methodological difficulties as label dependency and label imbalance further complicate the design of such systems.

Despite these challenges, MLC should not be considered impractical for GED. Several directions appear promising for future research. First, systematically analyzing the multilabel properties of learner corpora — e.g., through measures such as label cardinality or density — could help determine the extent to which GED data truly require multilabel modeling. Second, future work would benefit from explicitly reporting the classification paradigm adopted (for instance, by referring to our typology), in order to improve methodological transparency. Third, further exploring multilabel learning methods developed in other fields may provide effective alternatives to the multiclass frameworks currently dominant in GED.

More broadly, this work emphasizes the importance of aligning modeling choices with the structural properties of the data. When learner errors may belong to multiple categories simultaneously, ignoring this dimension may obscure certain error patterns and limit the interpretability of GED outputs. Addressing this gap therefore represents a promising direction to improve the analysis and modeling of errors produced by learners during foreign language learning.

[9] Generative LLMs are increasingly explored for this task of artificial data generation. For further discussion, see for example Li et al. (2024) and Yan et al. (2025).

5 Limitations

A first limitation of this study is that we do not systematically investigate the impact of language typology on the choice of classification paradigm. As noted in the introduction, inflectional and agglutinative languages seem more likely to accumulate multiple errors within a single token, making multilabel classification particularly relevant in these contexts. Nevertheless, systematic empirical comparisons would be needed to confirm this hypothesis.

Moreover, although we report the performance of models explicitly designed for MLC, we do not address the specific methodological challenges involved in evaluating multilabel classifiers as compared to multiclass ones.

Finally, our survey has focuses exclusively on MLC methods that have actually been applied in the GED literature. As a result, we have not discussed a number of alternative approaches which, although often more complex, could potentially yield higher performance.

References

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Julio Arroyo, Pietro Perona, and Elijah Cole. 2023. Understanding label bias in single positive multi-label learning. *Preprint*, arXiv:2305.15584.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. Leveraging LLMs for Bangla grammar error correction: Error categorization, synthetic data, and model evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8220–8239, Vienna, Austria. Association for Computational Linguistics.

Jasmin Bogatinovski, Ljupco Todorovski, Saso Dzeroski, and Dragi Kocev. 2021. Comprehensive comparative study of multi-label classification methods. *CoRR*, abs/2102.07113.

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.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Judith Casademont Moner and Elena Volodina. 2022. Swedish MuClaGED: A new dataset for grammatical error detection in Swedish. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 36–45, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Qiuyi Chen, Haiyang Zhang, Leqi Zhang, Changchun Li, Jia Wang, and Wei Wang. 2026. PU-BENCH: A UNIFIED BENCHMARK FOR RIGOROUS AND REPRODUCIBLE PU LEARNING. In *The Fourteenth International Conference on Learning Representations*.

Qianqian Cheng, Jun Huang, Huiyi Zhang, Sibao Chen, and Xiao Zheng. 2023. Improving multi-label learning by modeling local label and feature correlations. *Intelligent Data Analysis*, 27(2):379–398.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, and Omri Abend. 2020. Classifying syntactic errors in learner language. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 97–107, Online. Association for Computational Linguistics.

Leshem Choshen, Matanel Oren, Dmitry Nikolaev, and Omri Abend. 2021. SERRANT: a syntactic classifier for english grammatical error types. *CoRR*, abs/2104.02310.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. 2021. Multi-label learning from single positive labels. *CoRR*, abs/2106.09708.

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.

Joel Ehnroth and Yoonjoo Park. 2023. Correction of grammatical errors in swedish. *LU-CS-EX*.

Irene Fioravanti, Luciana Forti, and Stefania Spina. 2024. Automatic error detection: Comparing AI vs. human performance on L2 Italian texts. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 366–372, Pisa, Italy. CEUR Workshop Proceedings.

Nicolás E. García-Pedrajas, José M. Cuevas-Muñoz, Gonzalo Cerruela-García, and Aida de Haro-García. 2024. A thorough experimental comparison of multilabel methods for classification performance. *Pattern Recognition*, 151:110342.

Francisco Herrera, Francisco Charte, Antonio J. Rivera, and Maria J. del Jesus. 2016. *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Springer.

Shan Huang, Wenlong Hu, Bin Lu, Qiang Fan, Xinyao Xu, Xiaolei Zhou, and Hao Yan. 2024. Application of label correlation in multi-label classification: A survey. *Applied Sciences*, 14(19).

Shen Huang and Houfeng Wang. 2016. Bi-LSTM neural networks for Chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 148–154, Osaka, Japan. The COLING 2016 Organizing Committee.

Atakan Kara, Farrin Marouf Sofian, Andrew Bond, and Gözde Şahin. 2023. GECTurk: Grammatical error correction and detection dataset for Turkish. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 278–290, Nusa Dua, Bali. Association for Computational Linguistics.

Anisia Katinskaia, Maria Lebedeva, Jue Hou, and Roman Yangarber. 2022. Semi-automatically annotated learner corpus for Russian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 832–839, Marseille, France. European Language Resources Association.

Inga Kempfert and Christine Köhn. 2018. An automatic error tagger for german. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 32–40.

Katerina Korre, Marita Chatzipanagiotou, and John Pavlopoulos. 2021. ELERRANT: Automatic grammatical error type classification for Greek. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 708–717, Held Online. INCOMA Ltd.

Daniil Kosakin, Sergei Obiedkov, Ivan Smirnov, Ekaterina Rakhilina, Anastasia Vyrenkova, and Ekaterina Zalivina. 2024. Russian learner corpus: Towards error-cause annotation for L2 Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14240–14258, Torino, Italia. ELRA and ICCL.

Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. 2024. Data generation using large language models for text classification: An empirical case study. *Preprint*, arXiv:2407.12813.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Atsushi Mizumoto, Natsuko Shintani, Miyuki Sasaki, and Mark Feng Teng. 2024. Testing the viability of chatgpt as a companion in l2 writing accuracy assessment. *Research Methods in Applied Linguistics*, 3(2):100116.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

Regina Nasyrova and Alexey Sorokin. 2025. Grammatical error correction via sequence tagging for Russian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1036–1050, Vienna, Austria. Association for Computational Linguistics.

Li Nguyen, Shiva Taslimipoor, and Zheng Yuan. 2024. What can nlp do for linguistics? towards using grammatical error analysis to document non-standard english features. *Linguistics Vanguard*, 10(1):465–477.

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.

Laura Noreskal, Iris Eshkol-Taravella, and Marianne Desmets. 2023. Détecter une erreur dans les phrases coordonnées au sein des rédactions universitaires. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 248–261, Paris, France. ATALA.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop*

on Innovative Use of NLP for Building Educational Applications, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Nabilah Oshin, Syed Hoque, Md Fahim, Amin Ahsan Ali, M Ashraful Amin, and Akmmahbubur Rahman. 2023. BaTEClaCor: A novel dataset for Bangla text error classification and correction. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 124–135, Singapore. Association for Computational Linguistics.

Luca Piras, Ludovico Boratto, and Guilherme Ramos. 2021. Evaluating the prediction bias induced by label imbalance in multi-label classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3368–3372, New York, NY, USA. Association for Computing Machinery.

Ying Qin, Yumeng Luo, and Yuming Zhai. 2023. French error type annotation for dictation: A platform with automatic error type annotation for french dictation exercises. *Frontiers in Psychology*, Volume 13 - 2022.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1181–1191, Berlin, Germany. Association for Computational Linguistics.

Alla Rozovskaya. 2022. Automatic classification of Russian learner errors. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5637–5647, Marseille, France. European Language Resources Association.

Abdullah Shahid, Farah Kashif, Shazia Ayyaz, and Komal Ijaz. 2023. Teacher's implementation of error-correction strategies in l2 classroom. *Review of Applied Management and Social Sciences*.

Xinpei Su, Zhuojun Han, and Yitian Xu. 2026. Class-conditional unbiased loss for single-positive multi-label learning. *Applied Soft Computing*, 186:114064.

Harun Uz and Gülşen Eryiğit. 2023. Towards automatic grammatical error type classification for Turkish. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 134–142, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Verelst, Paul K. Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. 2022. Spatial consistency loss for training multi-label classifiers from single-label annotations. *Preprint*, arXiv:2203.06127.

Yixuan Wang, Yijun Liu, Bo Sun, and Wanxiang Che. 2023. System report for CCL23-eval task 8: Chinese grammar error detection and correction using multi-granularity information. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 271–281, Harbin, China. Chinese Information Processing Society of China.

Mengmeng Xu, Yancheng Bai, Bernard Ghanem, Boxiao Liu, Yan Gao, Nan Guo, Xiaochun Ye, Fang Wan, H You, D Fan, and 1 others. 2019. Missing labels in object detection. In *CVPR workshops*, volume 3.

Ning Xu, Congyu Qiao, Jiaqi Lv, Xin Geng, and Min-Ling Zhang. 2022. One positive label is sufficient: Single-positive multi-label learning with label enhancement. In *Advances in Neural Information Processing Systems*, volume 35, pages 21765–21776. Curran Associates, Inc.

Chunxiao Yan, Iris Eshkol-Taravella, Sarah De V ogué, and Marianne Desmets. 2025. Augmentation des données par LLM pour améliorer la détection automatique des erreurs de coordination. In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 154–166, Marseille, France. ATALA \\& ARIA.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean grammatical error correction: Datasets and annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. 2021. Re-labeling imagenet: from single to multi-labels, from global to localized labels. *CoRR*, abs/2101.05022.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. 2022. Acknowledging the unknown for multi-label learning with single positive labels. *Preprint*, arXiv:2203.16219.

A Appendix

Paper	Model type	Classification paradigm	Approaches
Fioravanti et al. (2024)	Generative LLM	Multiclass	Mutually exclusive typology
Choshen et al. (2020)	Symbolic	Multiclass	Mutually exclusive typology
Mizumoto et al. (2024)	Generative LLM	Multiclass	Without multilabel awareness
Bryant et al. (2017)	Symbolic	Multiclass	Without multilabel awareness
Náplava et al. (2022)	Symbolic	Multiclass	Without multilabel awareness
Katinskaia et al. (2022)	Symbolic	Multiclass	Without multilabel awareness
Uz and Eryiğit (2023)	Symbolic	Multiclass	Without multilabel awareness
Choshen et al. (2021)	Symbolic	Multiclass	Without multilabel awareness
Yuan et al. (2021)	Discriminative transformer	Multiclass	Without multilabel awareness
Nguyen et al. (2024)	Discriminative transformer	Multiclass	Without multilabel awareness
Kara et al. (2023)	Discriminative transformer and generative LLM	Multiclass	Without multilabel awareness
Bhattacharyya and Bhattacharya (2025)	Encoder–decoder	Multiclass	Without multilabel awareness
Korre et al. (2021)	Symbolic	Multiclass	Despite multilabel awareness
Yoon et al. (2023)	Symbolic	Multiclass	Despite multilabel awareness: label priority hierarchy
Kosakin et al. (2024)	Symbolic	Multiclass	Despite multilabel awareness: label priority hierarchy
Oshin et al. (2023)	Machine learning, deep learning and discriminative transformers models	Multiclass	Despite multilabel awareness: residual class for multiple errors
Rozovskaya (2022)	Symbolic	Partially multilabel	Multilabel only for morphological errors
Kempfert and Köhn (2018)	Symbolic	Partially multilabel	Multilabel only within the same grammatical class
Nasyrova and Sorokin (2025)	Discriminative transformer	Partially multilabel	Multilabel only for insertion labels on spaces between tokens
Qin et al. (2023)	Symbolic	Partially multilabel	Multilabel only between grammatical and phonetic errors
Belkebir and Habash (2021)	Symbolic	Multilabel	Algorithm adaptation
Noreskal et al. (2023)	Discriminative transformer	Multilabel	Algorithm adaptation
Yan et al. (2025)	Discriminative transformer	Multilabel	Algorithm adaptation
Huang and Wang (2016)	Deep learning model	Multilabel	Binary relevance
Wang et al. (2023)	Discriminative transformer	Multilabel	Binary relevance
Alhafni et al. (2023)	Discriminative transformer	Multilabel	Label powerset

Table 3: Overview of grammatical error classification approaches in GED, highlighting how different studies handle the multilabel nature of learner errors.