ANONYMOUS: A Novel Resource for English L2

Anonymous ACL submission

Abstract

The availability of suitable learner corpora is paramount for the study of second language acquisition (SLA) and language transfer. However, curating learner corpora is a challenging endeavor as high quality learner data is rarely publicly available. This results in only a few such corpora, such as ICLE and TOEFL-11, available to the community. To address this important gap, in this paper we present ANONYMOUS,[1] a novel English learner corpus with longitudinal data. ANONYMOUS contains texts written by adult learners taking English as a second language courses in the USA with the goal of either preparing for university admission or improving their language proficiency while starting their university degrees. ANONYMOUS contains 687 instances written by speakers of 15 different L1s. Unlike most learner corpora, the corpus contains longitudinal data which enables researchers to investigate language learning over time. We present two case studies using ANONYMOUS at the intersection of SLA and Computational Linguistics: (1) Native Language Identification (NLI); and (2) a quantitative and qualitative study using LLMs on linguistic features influenced by L1.[2]

1 Introduction

A language learners' first language (L1) often influences the fluency, grammatical patterns, and vocabulary usage in their second language (L2). This results in L2 production containing unique characteristics or linguistic features which may be unfamiliar or questionable to a native speaker.

Learner corpora can be used for a variety of tasks in SLA and Computational Linguistics. For example, the task of automatically identifying a language learners' first language based on these unique characteristics is known as Native Language Identification (NLI). NLI research has predominately focused on developing machine learning (ML) models to identify a learners' L1 through spoken features, such as pronunciation, stress, and prosodic patterns (Krishna et al., 2019). However, text-based NLI utilizes features such as word choices, syntax, and spelling to make predictions regarding an individual's mother tongue.

Less research has been conducted on text-based NLI, despite it having a number of use cases, including author profiling, forensics, spam and phishing detection, and a number of educational applications (Malmasi et al., 2017). In particular, this paper focuses on the use of NLI for identifying the native speaker of student essays. Various types of errors are then analyzed within each student essay to investigate whether a correlation can be drawn between a student's L1 and the type and frequency of spelling errors they produce. In turn, we demonstrate the ability of NLI to automatically recognize the first language of a student essay's author as well as its capacity to aid second language acquisition research.

The contributions of our work are the following:

1. We introduce ANONYMOUS, a novel corpus of L2 writing with longitudinal data. The corpus can be used for a variety of purposes in Computational Linguistics and SLA. We have IRB approval for this.

2. We describe the first linguistically-informed LLM-based study of features of L1 to L2 transfer on longitudinal data, also introducing more syntactically-informed analysis tools based on the concept of catenae.

3. We present various NLI experiments using this corpus. We evaluated the performance of various models, from traditional classifiers like SVMs, to state-of-the-art LLMs such as GPT-4.

[1] Anonymized to ensure double-blind review.

[2] All data and code will be made publicly available.

2 Related Work

2.1 Cross-linguistic Influence

Second Language Acquisition and Error Taxonomies A long tradition of SLA research has documented systematic learner errors commonly attributable to L1 interference (Richards, 1971; Odlin, 1989). While large learner corpora such as the Cambridge Learner Corpus (Nicholls, 2003) and NUCLE (Dahlmeier et al., 2013) include valuable metadata about each writer's L1, they do not typically annotate individual errors for cross-linguistic influence. Instead, error frameworks tend to focus on what is wrong; classifying the locus of the error (lexis, syntax, morphology) and the surface modification needed (e.g., omission, addition, substitution)—rather than *why* it emerged (Díaz-Negrillo and Fernández-Domínguez, 2006).

Grammatical Error Correction and LLMs Recent advancements in grammatical error correction (GEC) have followed from the emergence of large language models (LLMs) like GPT-3 and GPT-4, which have been evaluated for their performance in GEC tasks (Song et al., 2024; Kobayashi et al., 2024). For instance, studies have investigated the effectiveness of LLMs in GEC evaluation by employing prompts designed to incorporate various evaluation criteria (Loem et al., 2023; Fang et al., 2023). However, these models primarily focus on correcting errors rather than providing explanations that consider a learner's L1.

Recent research has attempted to go beyond grammatical error correction by considering L1 influences in academic writing. Zomer and Frankenberg-Garcia 2021 proposed a pre-trained encoder-decoder model designed to improve research writing by adapting corrections to the writer's L1 background. Their approach recognizes that L1 influences writing style and errors, offering targeted corrections based on linguistic transfer effects. However, their study primarily focuses on improving research writing rather than systematically analyzing or categorizing L1 interference at a linguistic level. Moreover, their model does not explicitly attribute errors to phonological, orthographic, or syntactic transfer from the L1.

Our Contribution In contrast to these approaches, our work is, to our knowledge, the first to use LLMs –paired with human oversight, of course– for explicit L1 interference analysis. Our prompt-driven annotation scheme goes beyond standard error detection by requiring the model to (1) identify whether an error stems from L1 interference and at what level (e.g., syntax, morphology) and (2) justify its label with concrete linguistic features from the learner's native language. By integrating SLA insights, we generate fine-grained annotations that capture L1 influence. This structured, L1-aware output moves beyond standard GEC tasks, bridging the gap between automatic correction and the deeper linguistic understanding emphasized in SLA research.

2.2 Native Language Identification

The underlying assumption in NLI is that the native language influences acquisition and production of second language, a phenomenon known as cross-linguistic influence or language transfer (Krashen, 1981; Ellis, 2015). Language transfer results in L1 features manifesting in L2 production, allowing computational models to recognize patterns shared by speakers of the same L1 when communicating in a given L2. Text-based NLI has numerous important applications such as serving as a corpus-driven approach for SLA (Jarvis and Crossley, 2012) and enabling the development of effective L2 teaching materials and computer-aided language learning (CALL) software. Additionally, NLI has been shown to improve NLP systems dealing with texts from non-native speakers, contributing to tasks like author profiling, forensics, spam and phishing detection (Malmasi et al., 2017).

As evidenced by a recent survey (Goswami et al., 2024), traditional statistical models such as Support Vector Machines (SVMs) trained on n-grams as features have historically delivered the best performance for text-based NLI. A few recent studies (Lotfi et al., 2020; Uluslu and Schneider, 2022; Zhang and Salle, 2023; Ng and Markov, 2025), however, have shown that fine-tuned LLMs such as GPT-4 deliver state-of-the-art performance for English NLI. In this paper, we test multiple approaches on this corpus capturing the full breadth of the available toolkit from including SVM ensembles all the way to the recently released GPT-4o.

3 The ANONYMOUS dataset

The ANONYMOUS dataset consists of student essays of various types provided by international students at a US R1 university. Students provided evidence specifying their country of origin and L1, yielding a sample of 687 essays written by students

			L1 Information		Annotations	
Dataset	**L1 Languages**	**Size**	**L1 Metadata**	**L1-Annotated Errors**	**Fine-Grained Errors**	**Longitudinal**
Cambridge Learner Corpus (CLC)	80+	~2.9M words	✓	✗	✓	✗
NUCLE (CoNLL-2014)	N/A	~1.9M words	✗	✗	✓	✗
FCE Corpus	European & Asian (16 L1s)	~1,200 essays	✗	✗	✓	✗
ICNALE	East/Southeast Asian (10 L1s)	~3.8M words	✓	? (some)	✓	✗
TOEFL11	Mixed (11 L1s)	~12,000 essays	✓	✗	✓	✗
EFCAMDAT	European & Asian (9 L1s)	~100K learners	? (nationality)	✗	✓	✓
BEA-2019 (W&I+LOCNESS)	N/A	334 learners	✗	✗	✓	✗
ANONYMOUS	Arabic, Chinese, Vietnamese, (plus Azerbaijani, Telugu, Dari, and 9 more)	57 (+26) learners	✓	✓	✓	✓

Table 1: Comparison of ANONYMOUS with other SLA learner corpora, highlighting L1 metadata and L1-marked errors. Dataset sources: CLC (Nicholls, 2003), NUCLE (Dahlmeier et al., 2013), FCE (Yannakoudakis et al., 2011), ICNALE (Ishikawa, 2023), TOEFL11 (Blanchard et al., 2013), EFCAMDAT (Geertzen et al., 2014), BEA-2019 (Bryant et al., 2019).

of 15 different L1s: Arabic, Azerbaijani, Chinese, Dari, French, Hindi, Indonesian, Korean, Kyrgyz, Pashto, Portuguese, Russian, Telugu, Urdu, Vietnamese.Prior to enrolling at the university, students were also required to show evidence of their L2 English proficiency. All students had obtained an International English Language Testing System (IELTS) score of 7 or higher or an equivalent proof of L2 English proficiency. Lastly, all students were at the post-graduate level having enrolled in a Masters-level course.

For the analysis and NLI experiments presented in this paper we use a sub-set of ANONYMOUS containing instances by speakers L1s which contained more than three students in the dataset. The goal is to provide enough individual variation so that models could capture features of L1s and not idiolects. This results in a sub-set with 525 instances written by 57 students of three different L1s: Arabic, Chinese, and Vietnamese. Table 2 depicts a summary of student demographics.

The essays come from an introductory course designed for students who are new to both the university and the U.S. Essays ranged from short question and answer style essays, whereby the student provides a one to two sentence answer to a given question, to longer ones spanning multiple paragraphs on a particular topic or group project. All of the essays pertained to students' experiences in the US and their adjustment to university life. Essays were submitted by the students electronically using the university's assignment portal. All essays have since been anonymized with each student's personal information being omitted. Examples of essays questions and student responses are provided in Table 3.

The dataset is annotated with various linguistic features in relation to spelling, lexical, grammatical and other types of L2 English learner errors, including those likely caused by L1 influence. These annotations include detailed classifications of misspellings, word usage errors, and syntactic mistakes, offering a comprehensive insight into error types and frequencies—valuable for both NLI and spelling error prediction. Fine-grained annotation was completed with the process discussed below, involving prompting `gpt-4o`. Two trained linguists reviewed a stratified sample of the annotations for accuracy and validity. Only those linguistic features agreed upon by both annotators as examples of the outlined categorizes are included in the final dataset, a decision to favor fewer but higher-quality annotations over potentially more noisy ones.

4 Dataset Analysis

For the rest of this work, we will focus our analysis on a sample of our dataset, encompassing the three L1 languages with the largest support in ANONYMOUS, namely Arabic, Chinese, and Vietnamese.

4.1 Fine-Grained L1-Interference Annotations

SLA-Grounded Annotation Framework We draw on established second language acquisition (SLA) research to develop an annotation framework for learner errors. The categories—phonetic misrepresentations, morphological overgeneralizations, and L1-based orthographic interference—reflect well-documented SLA phenomena, such as Spanish speakers inserting an "e" before /s/ clusters or the overextension of regular morphological rules, e.g., "buyed" for "bought" (Richards and Schmidt, 2011; Freeman et al., 2016; Kazazoğlu, 2020). Grounding our schema in SLA principles ensures theoretical and pedagogical relevance.

	Arabic	Chinese	Vietnamese
# Speakers	35	18	4
Avg. Age	25	32	27
Level of Education	post-grad	post-grad	post-grad
IELTS (English Proficiency)	7	7	7
# Total Assignments	345	133	47
- # Short Assignments	187	64	12
- # Long Assignments	158	69	35
- # Group Assignments	112	59	14

Table 2: Dataset statistics for our focus languages, including demographic information and number of assignments per type.

Assignment Type	Question	Student Answer
Short	Why are we asking you about the 'type of learning' that is happening at UNIVERSITY?	To know about what I get benefit from it.
Long	Dissertation Paper - Write about your experience at UNIVERSITY.	After few hours fly, two plant transfer finely I got to the destination, at the time I got to UNIVERSITY, I want through some test at the first floor of Goble Center, there was a professor come find me, she told me...
Group	Describe what you have learned from the group project.	The first, take away is that I can talk with me from the language activity is that most people have a perfect specking skill when it comes to home language. However, it changes when it comes to the second language that we are studying. Each person has their own skill that they are good at...

Table 3: Anonymized examples of question types and student answers taken from the ANONYMOUS dataset.

L1	Train	Dev	Test	Total
Arabic	275	35	35	345
Chinese	107	13	13	133
Vietnamese	37	5	5	47
Total	419	53	53	525

Table 4: Number of documents in the dataset split for model training and evaluation per L1.

Using LLMs for L1-Based Annotation Our key methodological contribution is leveraging LLMs to generate SLA-informed annotations at scale, significantly reducing the labor-intensive nature of traditional error annotation.

Conventional annotation processes require thousands of expert-annotator hours to construct large corpora, with estimates suggesting that annotating one million words could take 2000-5000 hours[3]. In contrast, our approach harnesses a prompt-driven large language model (LLM) to systematically classify errors, integrating SLA insights to provide structured, L1-aware annotations at scale. The prompt (see Appendix A) guides the model to:

- Identify each error's subcategory (orthographic, morphological, lexical, grammatical, etc.).
- Flag L1 interference when observed, referencing specific native-language forms (e.g., a Spanish "e+ s" cluster or Arabic morphological patterns).

LLM Annotations Align with SLA Patterns To ensure plausible cross-linguistic references, we undertook a multi-tiered verification process. One author spot-checked approximately 20% of the annotations, and a trained linguist examined the overall prompt design to confirm its linguistic soundness. We further shared a selection of model outputs with two additional linguists during collaborative review sessions. While not a full-scale, systematic audit,

[3] For context, manually annotating a corpus of this scale—similar to NUCLE (Dahlmeier et al., 2013)—at an estimated rate of 500 words per hour would require extensive expert labor. This estimate accounts for multiple annotation passes, as is standard in error correction corpora, and is derived from previous annotation efforts (Dahlmeier et al., 2013; Ng et al., 2014).

these steps helped validate that the model's attributions to learners' L1 features were generally coherent and aligned with SLA phenomena.

A notable outcome is that the model often produces **linguistically aligned annotations**—for instance, highlighting Chinese syntactic constructions or Arabic orthographic habits.

Figure 1a shows such examples of Chinese L1 interference, along with the explanation provided by the model, which has been verified by native bilingual speakers. In the first example, the student meant to say "in terms of X" but used the construction "in the X aspect" which is a direct translation of the Chinese phrase used for such references. In the second example, again the student directly translates a Chinese construction that stands for "at the time of X" which in English would be more appropriately conveyed as "at X" or "during X". Similarly, Figure 1b illustrates orthographic interference from Arabic L1 speakers. The example highlights a phonetic-based spelling error where the student writes "attande" instead of "attend." This error likely stems from the phonological differences between Arabic and English, particularly the absence of certain vowel representations in Arabic orthography. Since Arabic does not typically mark short vowels in its writing system, learners may inadvertently insert or alter vowel sounds when spelling English words. The model correctly attributes this error to phonetic interference, reflecting how Arabic speakers may rely on phonological approximations when encoding unfamiliar English word forms.

```
[
  {
    "incorrect": "in the learning aspect",
    "correct": "in terms of learning",
    "type": {
      "L1InterferenceSubcategory.SYNTACTIC_INTERFERENCE": 1
    },
    "l1_interference_reason": "Chinese syntax often uses
    phrases like '在...方面' which translates directly to 'in
    the... aspect', leading to syntactic interference."
  },
  {
    "incorrect": "at the time of UNIVERSITY",
    "correct": "at UNIVERSITY",
    "type": {
      "L1InterferenceSubcategory.SYNTACTIC_INTERFERENCE":
      0.8,
      "GrammaticalSubcategory.GRAMMATICAL": 0.2
    },
    "l1_interference_reason": "Direct translation of '在...的
    时候' from Chinese might lead to the use of 'at the time
    of', which does not align with the English syntactic
    structure."
  }
]
```

(a) Annotated learner errors illustrating L1 syntactic interference from Chinese.

```
[
  {
    "incorrect": "attande",
    "correct": "attend",
    "type": {
      "L1InterferenceSubcategory.ORTHOGRAPHIC_INTERFERENCE":
      0.7,
      "OrthographySubcategory.PHONETIC": 0.3
    },
    "l1_interference_reason": "Arabic speakers might add
    extra vowels or alter consonant sounds due to the absence
    of certain English phonemes in Arabic, leading to
    'attande' instead of 'attend'."
  }
]
```

(b) Annotated learner errors illustrating orthographic interference from Arabic.

Figure 1: Annotated learner errors illustrating interference from various L1s. Each entry includes the incorrect phrase, its corrected form, and an explanation of syntactic or orthographic influence.

4.2 Tracking Student Errors Over Time

We track student error patterns over time to analyze linguistic development and learning trajectories. Timestamped writing submissions enable longitudinal analysis at both individual and cohort levels. To ensure comparability across time periods, we normalize error counts against text length and assignment counts. This allows us to assess whether certain error types diminish with proficiency gains or persist, indicating deeper linguistic challenges. Of course, the expectation for an English proficiency course is that learner errors diminish over time.

None of the observed fluctuations (e.g., rising error counts in certain months, subsequent declines) reach statistical significance (see Appendix B). However, the fine-grained L1-based labels reveal that certain patterns persist—such as Arabic speakers' difficulties with vowel representation or literal syntactic translations from Chinese—suggesting that some cross-linguistic influences remain stable over time rather than disappearing with increased exposure to English (Odlin, 1989).

Our results seem to contradict our hypothesis that error frequencies should reduce – for the 2022 cohort, for instance, error frequencies largely increase from one assignment to the other until the last assignment! For 2024, the story is somewhat reversed. We plan to explore several possible explanations for these observations. For example, it might be that students do become better L2 speakers, but assignments also become harder, leading to more errors. Or, perhaps, it could be that the first assignment was by construction an easy one thus leading to fewer errors, and, if we discard it, for the

2023 and 2024 cohorts we might actually confirm our hypothesis that learner error frequencies reduce over time. We plan to explore these explanations finding more deeply in future work, also engaging with the instructors of the class as well as with the students themselves.

4.3 Lexical Development

Beyond tracking general error trends, we also explore lexical development in relation to Romance and Germanic vocabulary acquisition. Previous studies have documented that Germanic and Romance L1 speakers tend to overuse cognates from their respective L1s in English at lower proficiency levels, with this reliance decreasing as proficiency increases (Nativ et al., 2024). However, our focus dataset consists of Arabic, Chinese, and Vietnamese L1 speakers, for whom English lacks a strong lexical overlap with their native languages. Analyzing how these learners acquire vocabulary from different etymological sources represents a novel contribution to SLA research.

In theory, we expect to see an increasing tendency toward Romance-derived vocabulary as students advance in proficiency, given that academic and formal English draws heavily from Latin and French (Hernandez et al., 2021). Our analysis partially supports this: the 2022 cohort (see Figure 3) shows a statistically significant rise in Latin-based vocabulary over time (p = 0.0199). However, this trend vanishes in the 2023 and 2024 cohorts, raising questions about how learners from non-Indo-European backgrounds acquire academic vocabulary. Differences in instructional input, cognitive processing, or exposure to academic vocabulary may contribute to these variations. The observed increase in the 2022 cohort suggests that under certain conditions, learners do shift toward more Latin-derived vocabulary as they progress, highlighting the need for further research into the factors that influence this shift. Future studies should examine whether these trends persist across larger datasets and explore pedagogical interventions that could facilitate the acquisition of academic English vocabulary for learners from diverse linguistic backgrounds.

4.4 Further Syntactic Pattern Analysis

Syntactic analysis in NLP and second-language acquisition (SLA) research has traditionally relied on head-dependent relations within dependency trees (Constant et al., 2017). However, these relations often fail to capture multi-word syntactic units that function as a single structural unit. This is also the issue with analyses that focus on common Part-of-Speech n-grams.

Here, we propose to use *syntactic catenae* as the unit of analysis to remedy these issues. Osborne et al. (2012) introduced *catenae* as a more flexible syntactic representation, defining them as *any sequence of words that maintains a continuous dominance relationship in a dependency tree*. This definition allows catenae to include non-constituent structures and discontinuous elements that are crucial for syntactic analysis.

Catenae have been used in syntactic theory to describe verb complexes, idiomatic expressions, and discontinuous dependencies (Osborne et al., 2012; Imrényi, 2013). However, their application in corpus-based computational linguistics, particularly in L2 syntactic variation analysis, remains unexplored. We investigate whether catenae distributions exhibit L1-specific patterns in learner writing, exploring whether different L1 groups favor certain syntactic constructions when producing English.

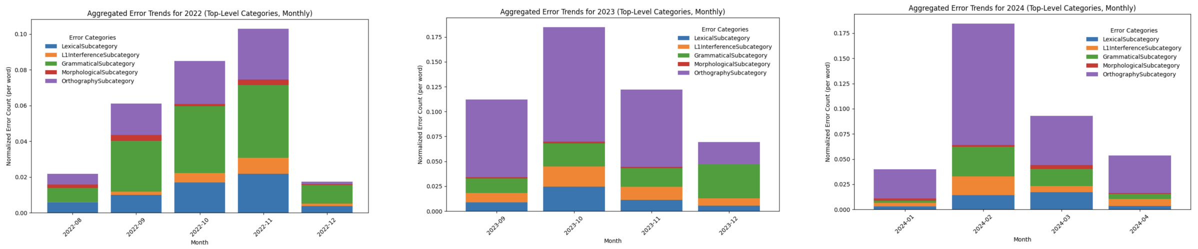
We additionally conduct a supplementary investigation using POS bigrams, which capture short-range syntactic dependencies (De Gregorio et al., 2024). While less structurally expressive than catenae, POS bigrams offer a more conventional means of detecting syntactic variation across L1 groups.

Methodology Using *Stanza* (Qi et al., 2020), we extract catenae from dependency-parsed texts, representing them as sequences of *(dependency relation, POS tag)* pairs (e.g., `det-DT | comp:obj-NN | mod-JJ`). This allows for a structural analysis independent of lexical choice. For interpretability, we also retain corresponding lexical sequences.

To supplement the catenae analysis, we also extract POS bigrams from learner texts, identifying adjacent POS sequences (e.g., `DT NN`, `NN VBZ`) as a proxy for syntactic tendencies across L1 groups.

Cross-L1 Comparison For both catenae and POS bigrams, we compute relative frequencies and apply TF-IDF weighting to identify structures that were more prominent in one L1 group relative to others.

Across both analyses, we do not observe *strong* L1-specific syntactic patterns. Frequent catenae were largely **shared across L1 groups**, with no consistent L1-driven structural tendencies. That said, we do observe some interesting differences



(a) 2022 Cohort

(b) 2023 Cohort

(c) 2024 Cohort

Figure 2: Aggregated Error Trends for Different Cohorts (Top-Level Categories, Monthly). The 2022 cohort shows a gradual increase in errors, peaking in November. The 2023 cohort exhibits higher orthographic errors throughout, while the 2024 cohort displays a sharp peak in February before declining.

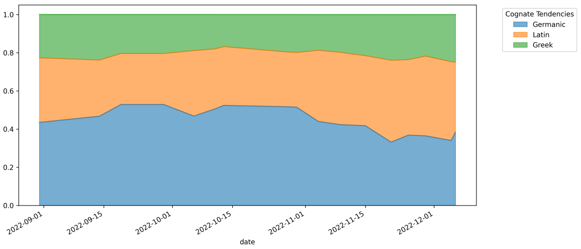


Figure 3: Proportion of Germanic, Latin, and Greek-derived vocabulary in learner writing over time (2022 cohort). The increase in Latin-based words suggests a shift toward academic vocabulary, while Germanic words remain dominant.

across different L1s. For example, compound noun constructions feature more prominently in Vietnamese L1 speakers and much less common in Chinese ones, even though one would probably expect the opposite due to the extensive compounding in Chinese.

Of course, we should note that the large space of possible catenae combinations and our rather sparse corpus limited our ability to detect robust differences. The relatively small number of speakers per L1 further constrained cross-L1 generalizability. We maintain, though, that catenae are the appropriate unit of analysis for uncovering L1-influenced syntactic patterns, and we leave such a larger scale analysis encompassing more corpora for future work.

5 Native Language Identification

As a further showcase of the utility of our dataset for other downstream tasks, we carry out multiple NLI experiments with results presented in Table 5. We present results in terms of accuracy and macro F1 score following the literature in this task (Goswami et al., 2024).

Statistical Ensemble We train multiple SVM systems using various features such as POS n-grams of $n \in [1, 4]$ and word n-grams of $n \in [1, 2]$. We then combine them in a majority voting ensemble (Malmasi and Dras, 2017) and we refer to this model as SVM Ensemble in the table.

BERT-based Models We fine-tune multiple BERT-based models on ANONYMOUS namely BERT, mBERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). We use learning rate of $1e-5$ for all models and early stopping on our development set.

LLMs We benchmark three LLMs on ANONYMOUS, namely FLAN-T5 (Chung et al., 2024), GPT-4o (Achiam et al., 2023), and the 70B parameter LLaMa 3.1 (Touvron et al., 2023). We benchmark the three models using both zero-shot prompting as well as task-specific fine-tuning on the training set.

NLI Takeaways Corroborating the results reported in recent studies using popular NLI datasets like TOEFL 11 (Ng and Markov, 2025), we observe that fine-tuned models achieve the highest performance on ANONYMOUS. All three LLMs obtain significant performance improvement from zero-shot prompting to task fine-tuning. The performance of LLMs using zero-shot prompting is, in turn, inferior to the performance of both SVM ensemble and the three BERT models. This indicates that off-the-shelf LLMs do not fare particularly well in identifying L1s without any specific task fine-tuning.

6 Conclusion and Future Work

We present ANONYMOUS, a first-of-its-kind dataset of learner English, which stands apart from others due to encompassing longitudinal data and

Approach	Models	Acc.	F1
Statistical			
	SVM Ensemble	0.75	0.73
BERT-based			
	roBERTa	0.79	0.75
	BERT	0.77	0.72
	mBERT	0.70	0.68
LLM Zero-shot			
	GPT 4o	0.66	0.66
	LLaMa3.1	0.41	0.43
	FLAN T5	0.32	0.37
LLM Fine-tunning			
	GPT 4o	0.97	0.96
	LLaMa3.1	0.87	0.84
	FLAN T5	0.66	0.53

Table 5: Results of different models on the ANONYMOUS dataset. LLMS require fine-tuning to outperform BERT-based and simple statistical approaches.

fine-grained L1 interference annotations. We showcase interesting analysis on three L1s, introduce new syntactic analysis units, and perform NLI experiments on a subset of our dataset.

Importantly, ANONYMOUS will continue expanding every year with each incoming student cohort. As a result, ANONYMOUS will facilitate exciting research directions in Second Language Acquisition research, while also presenting opportunities for challenging setups in the development of language learning applications.

7 Limitations

Our approach likely performs best for high-resource languages, as LLMs are trained predominantly on well-documented linguistic data. For low-resource languages with limited digital presence or sparse learner corpora, the model's ability to identify and explain L1 interference may be weaker, leading to noisier or less reliable annotations.

Additionally, while we conduct careful manual verification of a subset of model-generated annotations for the three L1s that we study in this paper, a more extensive validation process is likely needed to ensure consistency and reliability across diverse L1s.

A major challenge for the reproducibility of our work is the rapid evolution of LLMs (e.g., GPT-3.5, GPT-4), as results can depend on a specific model version that later might become unavailable. We chose to rely on the best currently available model to ensure higher quality annotations for our dataset, but future work could reproduce this effort with open-sourced/open-weight models to explore robustness to model variation. In addition, future work should evaluate performance across a broader range of linguistic backgrounds and explore strategies for maintaining reproducibility despite ongoing model updates.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:1–53.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

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.

J. De Gregorio, R. Toral, and D. Sánchez. 2024. Exploring language relations through syntactic distances and geographic proximity. *EPJ Data Science*, 13:61.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Ana Díaz-Negrillo and Jesús Fernández-Domínguez. 2006. Error Tagging Systems for Learner Corpora.

Revista española de lingüística aplicada, ISSN 0213-2028, Vol. 19, 2006, pags. 83-102, 19.

Rod Ellis. 2015. *Understanding second language acquisition 2nd edition*. Oxford university press.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Max R. Freeman, Henrike K. Blumenfeld, and Viorica Marian. 2016. Phonotactic constraints are activated across languages in bilinguals. *Frontiers in Psychology*, 7:702.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Efcamdat: The ef-cambridge open language database. In *Selected Proceedings of the 2012 Second Language Research Forum*, pages 240–254, Somerville, MA. Cascadilla Proceedings Project.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of NAACL*.

Arturo E. Hernandez, Juliana Ronderos, Jean Philippe Bodet, Hannah Claussenius-Kalman, My V. H. Nguyen, and Ferenc Bunta. 2021. German in childhood and latin in adolescence: On the bidialectal nature of lexical access in english. *Humanities and Social Sciences Communications*, 8(1):162.

András Imrényi. 2013. The syntax of Hungarian auxiliaries: A dependency grammar account. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 118–127, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge, London.

Scott Jarvis and Scott A Crossley. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detectionbased Approach*. Multilingual Matters.

Semin Kazazoğlu. 2020. The impact of l1 ınterference on foreign language writing: A contrastive error analysis. *Dil ve Dilbilimi Çalışmaları Dergisi*, 16(3):1168–1188.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Stephen Krashen. 1981. Second language acquisition. *Second Language Learning*.

G Radha Krishna, R Krishnan, and VK Mittal. 2019. An automated system for regional nativity identification of indian speakers from english speech. In *Proceedings of IEEE INDICON*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of COLING*.

Shervin Malmasi and Mark Dras. 2017. Multilingual native language identification. *Natural Language Engineering*.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of BEA*.

Liat Nativ, Yuval Nov, Noam Ordan, Shuly Wintner, and Anat Prior. 2024. Do more proficient writers use fewer cognates in l2? a computational approach. *Bilingualism: Language and Cognition*, 27(1):84–94.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, Abu Dhabi, UAE. 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, page 572–581. Cambridge University Press Cambridge.

Terence Odlin. 1989. *Language Transfer*. Cambridge Applied Linguistics. Cambridge University Press.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Jack C. Richards. 1971. A non-contrastive approach to error analysis1. *ELT Journal*, XXV(3):204–219.

Jack C. Richards and Richard W. Schmidt. 2011. *Longman Dictionary of Language Teaching and Applied Linguistics*, 4th edition. Routledge, London.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Yavuz Uluslu and Gerold Schneider. 2022. Scaling native language identification with transformer adapters. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 298–302, Trento, Italy. Association for Computational Linguistics.

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.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Gustavo Zomer and Ana Frankenberg-Garcia. 2021. Beyond grammatical error correction: Improving L1-influenced research writing in English using pretrained encoder-decoder models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2534–2540, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Prompt

LLM Annotation Prompt

Task: You are an expert at identifying and classifying spelling and language errors made by English learners. Your highest priority is to identify errors that may be due to L1 (native language) interference and provide a brief but **specific** explanation of how the L1 could cause such an error. Your explanation should include:

- A **concrete linguistic example** from the L1 (e.g., a word or phrase in the learner's native language) or a well-known phonological, orthographic, or syntactic feature of the L1 that contributes to the error.
- A short discussion of how that L1 feature leads the learner to produce the erroneous English form.

If there is **no** L1 interference, classify the error into one of the following categories: orthographic, lexical, morphological, grammatical, or typo.

Steps to follow for each erroneous word:

1. **Determine if L1 interference is involved.**
    - If **yes**, select the appropriate L1 interference subcategory and provide a `"l1_interference_reason"` that:
        - Identifies the specific L1 feature (e.g., a Spanish prefix rule, an Arabic root pattern, a Japanese phonological constraint).
        - Explains how that feature maps to the incorrect English form.
    - If **no**, classify under other subcategories: orthographic, lexical, morphological, grammatical, or typo.
2. **Return the errors in the order they appear in the text.**

Error Categories and Descriptions

1. **Orthography Subcategories**
    - **Phonetic Errors**
        - Definition: Words spelled purely by sound, ignoring English orthographic norms.
        - Examples:
            - *fone → phone*
            - *nife → knife*
    - **Vowel Substitution and Omission**
        - Definition: Substituting or omitting vowels incorrectly.
        - Examples:
            - *hop → hope*
            - *beter → better*
    - **Silent Letters and Irregular Spelling**
        - Definition: Ignoring or mishandling silent letters or irregular spelling patterns.
        - Examples:
            - *clim → climb*
            - *writting → writing*
    - **Consonant Substitution Errors**

    - Definition: Replacing one consonant with another.
    - Examples:
      - *shose → chose*
      - *joke → yoke*
  - **Hyphenation, Compound Words, and Spacing Errors**
    - Definition: Errors in spacing or hyphenation of compound words.
    - Examples:
      - *infact → in fact*
      - *some where → somewhere*

2. **Lexical Subcategories**

   - **Homophone Confusion**
     - Definition: Mixing up words that sound alike but differ in spelling and meaning.
     - Examples:
       - *their → there*
       - *peace → piece*
   - **Lexical Errors**
     - Definition: Errors involving incorrect word choice due to misunderstanding of meaning.
     - Examples:
       - *among → below*
       - *borrow → lend*
   - **Phonological Confusion**
     - Definition: Errors where words are confused due to phonological similarities, often involving metathesis, substitution of similar phonemes, or confusion between near-homophones.
     - Examples:
       - *aboard → abroad* (Metathesis: reversed phonemes)
       - *form → from* (Transposition of adjacent sounds)
       - *claps → class* (Substitution of "p" for "s")

3. **Morphological Subcategories**

   - **Morphemic Errors with Affixes**
     - Definition: Incorrect handling of prefixes or suffixes.
     - Examples:
       - *beautifull → beautiful*
       - *hoping → hopping*
   - **Overgeneralization of Spelling Rules**
     - Definition: Applying English morphological or spelling rules too broadly.
     - Examples:
       - *buyed → bought*
       - *goed → went*

4. **L1 Interference Subcategories**

   - **Orthographic Interference**
     - Definition: Applying L1 spelling conventions to English.
     - Examples:

        - *esplendid* → *splendid* (Spanish: adding "e" before "s" clusters)
        - *colur* → *colour* (British vs. American orthography confusion)
- **Lexical Interference**
    - Definition: Using L1-based lexical forms or cognates in English.
    - Examples:
        - *telefon* → *telephone* (Spanish or German influence)
        - *facilitate* → *faciliter* (French influence)
- **Grammatical Interference**
    - Definition: Applying L1 grammatical patterns to English.
    - Examples:
        - *She has 24 years* → *She is 24 years old* (Spanish: "Ella tiene 24 años")
        - *He doesn't know nothing* → *He doesn't know anything* (Negative concord in some L1s)
- **Syntactic Interference**
    - Definition: Applying L1 syntactic structures to English.
    - Examples:
        - *He to the store goes* → *He goes to the store* (German word order influence)
        - *Beautiful is she* → *She is beautiful* (Japanese syntax influence)

5. **Grammatical Subcategories**

    - **Grammatical Errors**
        - Definition: Errors in grammar, syntax, word order, or agreement.
        - Examples:
            - *She go yesterday* → *She went yesterday*
            - *He like apples* → *He likes apples*

Categories and Subcategories:
We define a hierarchical categorization system using Python enums for clarity and consistency:

```python
from enum import Enum

class OrthographySubcategory(Enum):
    PHONETIC = "Phonetic Errors"
    VOWEL_SUBSTITUTION_OMISSION = "Vowel Substitution and Omission"
    SILENT_LETTERS_IRREGULAR = "Silent Letters and Irregular Spelling"
    CONSONANT_SUBSTITUTION = "Consonant Substitution Errors"
    HYPHENATION_SPACING = "Hyphenation, Compound Words, and Spacing Errors"
    CONSONANT_DOUBLING = "Consonant Doubling and Dropping"
    CAPITALIZATION_PUNCTUATION = "Capitalization and Punctuation Errors"
    TYPO = "Typo"

class LexicalSubcategory(Enum):
    HOMOPHONE_CONFUSION = "Homophone Confusion"
    LEXICAL = "Lexical Errors"
    PHONOLOGICAL_CONFUSION = "Phonological Confusion"

class MorphologicalSubcategory(Enum):
    MORPHEMIC_AFFIX = "Morphemic Errors with Affixes"
    OVERGENERALIZATION = "Overgeneralization of Spelling Rules"
    CONSONANT_DOUBLING = "Morphological Consonant Doubling and Dropping"

class L1InterferenceSubcategory(Enum):
    ORTHOGRAPHIC_INTERFERENCE = "Orthographic Interference"
    LEXICAL_INTERFERENCE = "Lexical Interference"
    GRAMMATICAL_INTERFERENCE = "Grammatical Interference"
    SYNTACTIC_INTERFERENCE = "Syntactic Interference"

class GrammaticalSubcategory(Enum):
    GRAMMATICAL = "Grammatical Errors"
```

Probabilities:

- For each error, provide a `"type"` field as an object where keys are the enum names (e.g., `"OrthographySubcategory.PHONETIC"`) and values are probabilities (floats).
- Probabilities must sum to `1.0` for that error.

If L1 Interference is detected:

- Include `"l1_interference_reason"` explaining how the L1 caused the error.

Output Format:
Return a JSON array of objects. Each object should contain:

- `"incorrect"`: the misspelled or erroneous word.
- `"correct"`: the correct form.
- `"type"`: a dictionary of `{error_type: probability}` where probabilities sum to 1.0.
- `"l1_interference_reason"`: a string if L1 Interference applies.

Format strictly as JSON, with no additional commentary.

Few-Shot Examples:

Example Input:

> **L1: Spanish**
> **Text:** After the long *fly* and waiting two hours, I saw a *plant* arrive, which I thought was the right one because it looked so *esplendid* even though I felt *beter* knowing I had finally gotten there. The *clim* was tough, but I *buyed* a ticket, carrying my *childs* with rain, my friend said he'd *shose* a seat for me, but *infact* issues we had. *im* sad.

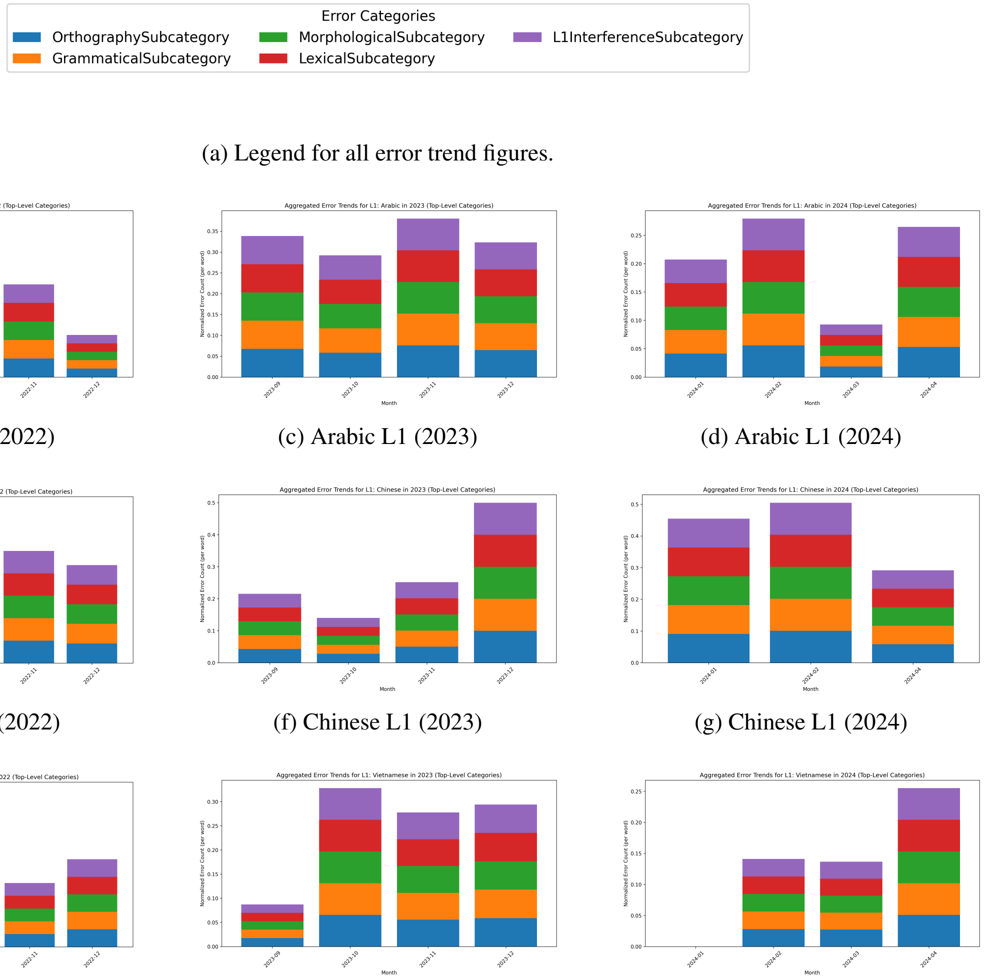
Example Output:

```
[
  {
    "incorrect": "plant",
    "correct": "plane",
    "type": {
      "OrthographySubcategory.PHONETIC": 0.8,
      "OrthographySubcategory.CONSONANT_SUBSTITUTION": 0.2
    }
  },
  {
    "incorrect": "esplendid",
    "correct": "splendid",
    "type": {
      "L1InterferenceSubcategory.ORTHOGRAPHIC_INTERFERENCE": 0.7,
      "OrthographySubcategory.PHONETIC": 0.3
    },
    "l1_interference_reason": "Spanish speakers often add an 'e' before 's' clusters due to L1
         orthographic habits."
  },
  ...
]
```

Note: This is a truncated example. The full prompt can be found the github repo.

B Error Trends by L1 and Year

In this section, we present the aggregated error trends for each L1 group across different years. Each plot shows the distribution of top-level error categories normalized by text length.



(a) Legend for all error trend figures.

(b) Arabic L1 (2022)

(c) Arabic L1 (2023)

(d) Arabic L1 (2024)

(e) Chinese L1 (2022)

(f) Chinese L1 (2023)

(g) Chinese L1 (2024)

(h) Vietnamese L1 (2022)

(i) Vietnamese L1 (2023)

(j) Vietnamese L1 (2024)

Figure 4: Aggregated error trends by L1 and year. Each subfigure represents a different L1-year combination.