Tracing L1 Interference in English Learner Writing Using a Longitudinal Corpus with Error Annotations

Poorvi Acharya, J. Elizabeth Liebl, Dhiman Goswami
Kai North, Marcos Zampieri, Antonios Anastasopoulos
George Mason University
{pachary4, antonis}@gmu.edu

Abstract

Language transfer is an important topic of research in second language acquisition and computational linguistics. 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 available to the community. To address this important gap, in this paper we present LENS, a novel English learner corpus with longitudinal data which enables researchers to investigate language learning over time. LENS contains 687 instances written by speakers of 15 different L1s. We use LENS two perform two important tasks at the intersection of SLA and Computational Linguistics: (1) Native Language Identification (NLI); and (2) an evaluation of large language models as a tool for high-precision, semi-automated annotation of L1 interference features.[1]

1 Introduction

A language learner's native language (L1, or first language) often influences fluency, grammatical patterns, and vocabulary usage in their second language (L2) (Ortega, 2014; Gass et al., 2020). This influence can result in L2 production containing distinctive linguistic features that often differs from native speakers' production.

These features have been the focus of research in Second Language Acquisition (SLA) and Computational Linguistics (CL), particularly through the use of learner corpora. These corpora enable the systematic analysis of language learner production and support key tasks, including automated proficiency assessment (Yannakoudakis et al., 2011; Vajjala and Rama, 2018) and Native Language Identification (NLI) (Malmasi et al., 2016). NLI involves automatically identifying a learner's L1 based on linguistic patterns in their L2 writing or speech. An important part of NLI research focuses on developing machine learning (ML) models for spoken language, analyzing features such as pronunciation, stress, and prosodic patterns (Krishna et al., 2019). Multiple studies have also addressed text-based NLI, which leverages written features, including word choice, syntax, and spelling, to make predictions about an individual's native language (Goswami et al., 2024).

Text-based NLI has a wide range of applications, including author profiling, forensics, spam and phishing detection, and various educational uses (Malmasi et al., 2017). Apart from a few notable exceptions (Ng and Markov, 2025; Goswami et al., 2025), however, the task has received limited attention in recent years. This paper investigates the use of NLI to identify authors' native languages in student essays. Each student essay is analyzed for various error types to explore potential correlations between the student's L1 and the types and frequencies of errors produced. We demonstrate how NLI can be used not only to automatically identify an author's L1, but also to contribute to research SLA.

The contributions of our work are the following:

1. We introduce the **L**ongitudinal **E**nglish **N**on-native **S**peaker (LENS) corpus. LENS is a novel corpus of L2 writing with longitudinal data, designed for a range of applications in CL and SLA.
2. We describe the first linguistically-informed LLM-based study of features of L1-to-L2 transfer on longitudinal data.
3. We present a series of NLI experiments using this corpus and evaluate the performance of models ranging from traditional classifiers, such as SVMs, to state-of-the-art large language models (LLMs), including GPT-4.

[1] https://github.com/p-acharya/LENSCorpus

2 Related Work

2.1 SLA and Error Taxonomies

A substantial body of SLA research has documented systematic learner errors often attributed 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 the nature of the error by classifying its locus (lexis, syntax, morphology) and the type of surface modification required (e.g., omission, addition, substitution), rather than investigating *why* it emerged (Díaz-Negrillo and Fernández-Domínguez, 2006).

2.2 Grammatical Error Correction and LLMs

Recent advancements in learner writing analysis have been driven by the emergence of LLMs. One major line of research focuses on Grammatical Error Correction (GEC), where models like GPT-4 are prompted to correct errors (Song et al., 2024; Kobayashi et al., 2024; Loem et al., 2023; Fang et al., 2023). In a parallel line of research, LLMs have been used for holistic assessment, such as replicating TOEFL11 essay scores (Mizumoto and Eguchi, 2023) or predicting CEFR proficiency levels (Benedetto et al., 2024). However, common to both of these approaches is a focus on the output — either a corrected text or an overall score — rather than the diagnostic process in the context of a learner's L1. These models primarily answer "Is this correct?" or "How good is this?", but are not designed to explain *why* a specific error was made.

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, the study primarily focuses on enhancing research writing rather than systematically analyzing or categorizing L1 interference at a linguistic level, and the model does not explicitly attribute errors to specific sources of transfer, such as phonological, orthographic, or syntactic influences from the L1.

2.3 Computational SLA Modeling

While our work focuses on the analysis of learner-produced text, another branch of Computational Second Language Acquisition (SLA) modeling aims to simulate the cognitive processes of learning. A study by Settles et al. (2018), for example, used large-scale data from a language-learning application to create statistical "half-life regression" models that predict memory decay for vocabulary items. Their model incorporates the learner's L1 background as a predictive feature, learning distinct forgetting curves for learners from different native languages. More recently, Stearns et al. (2024) have focused on evaluating the cognitive plausibility of "artificial learners" (neural networks) by testing their ability to generalize linguistic rules to unseen contexts.

The findings from these process-oriented models suggest that L1 background can be a factor in learning trajectories. Our research contributes not by modeling the learner's internal cognitive state, but by providing a corpus of learner *output* with fine-grained annotations of L1 interference. Such a dataset provides observable, micro-level linguistic data that may help explain the macro-level predictive effect of the L1 feature in cognitive models like those of Settles et al. (2018).

Our Contribution In contrast to these approaches, our work is, to our knowledge, the first to use LLMs paired with human oversight for explicit L1 interference analysis. We require the model to identify whether an error stems from L1 interference and at what level (e.g., syntax, morphology) and justify the 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 goes *beyond* standard GEC tasks, helping to bridge the gap between automatic correction and the emphasis on deeper linguistic analysis in SLA research.

2.4 Native Language Identification

NLI operates on the assumption that a learner's native language shapes the acquisition and production of a second language, a phenomenon referred to 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

Dataset	L1 Languages	Size (words)	L1 Information		Annotations	
			L1 Metadata	L1-Annotated Errors	Fine-Grained Errors	Longitudinal
Cambridge Learner Corpus (CLC)	80+	16M	✓	✗	✓	✗
NUCLE (CoNLL-2014)	—	1.22M	✗	✗	✓	✗
FCE Corpus	16 (EU/Asia)	532K	✗	✗	✓	✗
ICNALE	10 (E/SE Asia)	4.02M	✓	?	✓	✗
TOEFL11	11	4.21M	✓	✗	✓	✗
EFCAMDAT	9	83.5M	?	✗	✓	✓
BEA-2019 (W&I+LOCNESS)	—	715K	✗	✗	✓	✗
LONGDALE	—	780K	✓	✗	✗	✓
LENS	Arabic, Chinese, Vietnamese (+12 others)	175K	✓	✓	✓	✓

Table 1: Comparison of LENS with widely used SLA learner corpora. 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), LONGDALE (Meunier, 2016). ✓ = present; ✗ = absent; ? = partial/indirect.

communicating in a given L2. Text-based NLI has a number of 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 when 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 (Gebre et al., 2013; Goutte and Léger, 2017; Zampieri et al., 2017). 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 LENS Corpus

Collection context LENS was gathered from an introductory academic-writing course taken by international post-graduate students at a U.S. R1 university between 2022 and 2024.Learners produced three assignment types (short answers, long essays, and group reflections) based on their university and U.S. acculturation experiences, submitting work electronically via the learning-management system.

Learner profile Students self-reported their L1 and country of origin upon enrollment; no participant listed multiple L1s. All had demonstrated advanced English proficiency (IELTS ≥ 7) and were enrolled in Master's programs. The full corpus contains a total of 687 essays from 15 L1s.[2] Because many L1s are represented by only one or two learners, all analyses in this paper focus on the three languages with at least three writers: Arabic, Chinese, and Vietnamese.

Corpus Composition and Per-L1 Breakdown LENS includes a subset comprising texts from Arabic, Chinese, and Vietnamese learners. Table 2 presents the overall size of the subset analyzed in this paper, including the per-L1 breakdown, which details the number of learners, documents, tokens, median document length, mean submissions per learner, and the median number of weeks between first and last submission (a proxy for longitudinal depth).

Note that the three cohorts differ in how often they submitted short versus long tasks (Table 2). A detailed breakdown of the corpus's longitudinal properties, including submission statistics per cohort, can be found in Appendix Table 12.

Positioning among existing learner corpora Table 1 contrasts LENS with the most frequently used English-learner resources[3]. As shown in Table 1, while LENS is smaller in token count than large-scale resources like NUCLE, its unique contribution lies in being the first corpus, to our knowledge, to combine four key features: fine-grained error labels, explicit L1 transfer annotations, detailed per-

[2]The full list of L1s is in Appendix Table 11.

[3]For a comprehensive, community-maintained list of learner corpora, see the University of Louvain's "Learner Corpora Around the World" reference table (Centre for English Corpus Linguistics (CECL), 2024).

L1	Learners	Docs	Tokens	Median tok/doc	Entries per learner	Span(wks)	Count		Proportion	
							Long	Short	Long	Short
Arabic	35	345	63,090	79	9.86	10	158	187	0.47	0.53
Chinese	18	133	28,835	88	7.39	4	69	64	0.50	0.50
Vietnamese	4	47	12,471	199	11.75	11.9	35	12	0.70	0.30
Total	57	525	104,396	-	-	-	262	263	-	-

Table 2: Corpus subset composition and per-L1 breakdown, including the total number of documents, tokens, learners, and document types analyzed in this paper. A full breakdown of the corpus composition can be found in Appendix D, Table 11.

learner metadata, and multiple submissions over time."

This unique combination enables research questions that have been difficult to pursue with previous datasets, such as modeling the trajectory of cross-linguistic influence throughout an academic term (§4) or leveraging L1-aware error signals for few-shot native-language identification (§6).

Examples and splits Table 4 shows anonymized excerpts of each assignment type, while Table 3 gives the training, development, and test split used in our experiments.

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 3: Document counts in the train/dev/test split.

4 Error Annotations

SLA-Grounded Annotation Scheme We draw on established research in Second Language Acquisition (SLA) to develop an annotation framework for learner errors, organized into five domains: orthographic, lexical, morphological, grammatical, and L1 interference. Each domain is further subdivided into more fine-grained categories, which can be found in full in Appendix A. Importantly, some errors, particularly those involving L1 interference, may be classified under multiple domains. For instance, when a Vietnamese L1 learner produces experiment instead of experience, the error constitutes both a lexical error (inappropriate word choice) and an instance of L1 interference, given that the Vietnamese word th nghim may translate into either experiment or experience. The categories in our taxonomy were designed to reflect well-documented SLA phenomena, and all manually validated errors in our corpus fall within at least one of these domains.

Our framework is informed by contrastive analysis research in SLA, which highlights the influence of L1 transfer on L2 acquisition (Saeed Al-Sobhi, 2019). It also draws on earlier error coding schemes, such as those proposed by (de Haan, 2000) and (**?**). While our categories are broadly aligned with those of de Haan, sharing the principle that errors may belong to multiple classes, Nicholls' scheme was ultimately too fine-grained for our purposes, as it prioritizes the nature of the correction over the underlying source of the error.

Following Lüdeling and Hirschmann (2015), we recognize that the goal of error annotation is not to establish a universal tagset for the field, but to design a taxonomy tailored to the specific phenomena under investigation. We use a broader categorization for L1 errors, with the expectation that the key mistake will be specified in the L1 interference explanation string, rather than introducing an excessive number of categories. This approach remains grounded in SLA principles, maintaining both theoretical and pedagogical relevance.

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 [4].

Conventional annotation processes require thousands of expert-annotator hours to construct large corpora, with estimates suggesting that annotating one million words could take 2,000-5,000 hours[5].

In contrast, our approach harnesses a prompt-

[4]All experiments were conducted using the gpt-4o-2024-08-06 model version accessed between Dec 11th 2024 and Jan 7th 2025, with a temperature setting of 0.7.

[5]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).

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...
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...

Table 4: Anonymized examples of the three assignment types.

driven 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).

We then extract the exact error span. Figure 1a shows examples of Chinese and Arabic L1 interference, verified by native bilingual speakers.

```
{
  "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.",
  "span_start": 3136,
  "span_end": 3158
}
```

(a) Annotated learner errors illustrating syntactic interference from Chinese L1, where direct translations of native constructions result in non-standard English expressions.

```
{
  "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 L1, where phonetic spelling errors arise from the lack of vowel marking in Arabic.

Figure 1: Each entry contains the incorrect phrase, its span, the corrected form, and an explanation of the interference type.

4.1 Modeling Error Rate Differences Across Assignment Types

To account for repeated submissions, we fit a Poisson Generalized Estimating Equations (GEE) model, which revealed that short answers exhibit a significantly higher error rate than long essays (β=1.24, p<0.001). A qualitative follow-up analysis suggests this disparity is driven by the *type* of errors common to each format. As shown in Table 8, short-form answers contained a higher proportion of surface-level orthographic and morphological errors, which the LLM detects reliably. In contrast, longer essays featured more complex syntactic and lexical choice errors, which the model is more likely to miss. This indicates that the observed error rate difference is partly an artifact of the LLM systematically under-reporting errors in longer submissions, a key limitation when comparing error rates across texts of varying lengths. This pattern also aligns with SLA research suggesting that task constraints influence error types.

4.2 Detecting keyboard typos

To determine whether the LLM occasionally assigns high-stakes labels to errors that are really just keyboard slips, we compared the QWERTY keyboard-distance distribution of the Typo category with every other sub-category using Welch's t-test (Table 9)[6].

The keyboard-distance analysis suggests that the LLM is generally well-calibrated for high-level categories (e.g., Grammatical, Lexical, L1-Interference). However, it tends to over-label several low-level orthographic phenomena.

In particular, consonant-doubling, consonant-substitution, morphological, and phonetic errors

[6] Our analysis is based on the heuristic of a standard QWERTY layout. For a more fine-grained analysis of typing behaviors that account for linguistic properties beyond simple keyboard distance, see recent work such as Pacquetet (2024) and Velentzas et al. (2024).

often resemble typos in terms of key proximity. Statistical analysis revealed no significant difference in mean distances for these error types when compared to genuine typos ($p > .05$). This indicates that many such tokens could be re-classified as benign slips rather than systematic errors. In contrast, errors with significantly larger mean distances than typos ($|t| \geq 2.08$, $p < .05$) include grammatical, lexical, L1 interference, hyphenation / spacing, silent-letter / irregular, and vowel substitution / omission. These categories typically involve changes that go beyond adjacent-key slips, suggesting a more substantive error rather than a mere typo. Interestingly, capitalization/punctuation and the broader punctuation class showed smaller average distances compared to typos (t= 2.49 and 3.11; p=.017 and .004). This pattern is consistent with same-key mistakes, such as missed shift keys, rather than cross-key substitutions.

These findings motivate two main adjustments: (i) implementing a post-processing rule to downgrade low-distance instances within borderline subcategories, and (ii) refining prompt engineering to explicitly consider keyboard proximity when distinguishing between typos and more substantial errors.

However, we do not remove labels for errors that resemble typos solely based on keyboard proximity. The fact that some morphological or phonetic errors have similar distances to genuine typos does not imply they are typographical mistakes; such errors may still arise from systematic L1 interference or language processing challenges. Therefore, we interpret the similarity as a potential confounding factor rather than grounds for exclusion.

4.3 Human Verification of GPT-4 Annotations

To assess the reliability of our automatically-generated labels, we employed a two-tier human-in-the-loop verification process. This approach combines document-level recall checks with native-speaker scrutiny of L1 interference claims, providing a principled estimate of annotation quality.

Verification Process All essays and error snippets were presented in a web interface that allowed span-level confirmation or correction; corrections were stored as an additional layer in the corpus. Disagreements were discussed in weekly meetings to ensure consistent annotation practices. We emphasize that this two-stage process was designed to provide a principled estimate of annotation quality, rather than to serve as a formal inter-annotator agreement study, which we leave for future work. The verification process involved two stages:

- **L1-Specific Check:** Two native-speaker linguists (Arabic and Mandarin) independently evaluated 10 randomly-selected errors per language flagged as L1 interference by GPT-4. They answered the following questions:
  - **Q1 (Plausibility):** Is this a plausible case of L1 interference? (Yes / No + rationale).
  - **Q2 (Explanation):** Is GPT-4's explanation of the interference accurate? (Yes / No + rationale).

  Native-speaker acceptance rates were 100% for both Arabic and Mandarin.
- **Document-Level Audit:** A third linguist, experienced in corpus annotation, audited 13% of the essays (stratified by L1 and assignment type). The linguist evaluated whether:
  - GPT-4 correctly identified errors or missed any errors.
  - Identified errors were correctly typed (orthographic, morphological, grammatical, etc.).
  - For errors labeled as L1 interference, both the attribution and the explanation were accurate.

4.4 Evaluation

Table 5 presents the precision, recall, and F1 score for each evaluation aspect.

Metric	Precision	Recall	F1 Score
Error Detection	0.916	0.107	0.191
Correction Agreement	0.697	0.083	0.149
Type Agreement	0.613	0.074	0.132
L1 Reason Agreement	0.837	0.038	0.072

Table 5: Overall performance metrics for LLM annotations compared to human annotations.

Interpretation Our interpretation is that GPT-4o is not suitable for fully automated annotation due to its low overall recall (11%), but is highly effective as a *first-pass annotation tool* in a semi-automated pipeline. Its high general precision (92%) means that when the model flags an issue, it is usually correct, making it a reliable assistant for human annotators. The model performs well on clear, surface-level issues like typos and lexical interference (see Table 10), but its true strength emerges in the more challenging task of identifying L1 interference.

This primary challenge—the automated identification of L1 interference—is where the model's high-precision, low-recall profile proves most valuable. While the model often misses subtle instances of L1 transfer, its key strength is the exceptional reliability of the errors it does find. In our L1-specific verification, native-speaker linguists unanimously confirmed (100%) the plausibility of every L1 interference case flagged by the model for both Arabic and Mandarin.

This result highlights the model's ability to serve as a high-confidence discovery tool for L1 transfer. This performance profile is ideally suited for a semi-automated workflow focused on building corpora of L1 errors, where the model's suggestions can be trusted as high-quality candidates for human review. Overall, while future improvements should target enhancing recall to create a more comprehensive tool, these results establish GPT-4o as an effective specialist for the high-precision discovery of L1-influenced errors in learner writing.

	LLM	
Annotator	Error	NotError
Error	113	912
NotError	7	0

Table 6: Confusion matrix (correct,wrong) for error detection between LLM and human annotator. See Appendix Figure 4 for detailed confusion matrix by error type.

5 Data Analysis

5.1 Tracking Student Errors Over Time

As timestamped writing submissions enable longitudinal analysis at both individual and cohort levels, we track student error patterns over time to analyze student development and learning trajectories. 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.

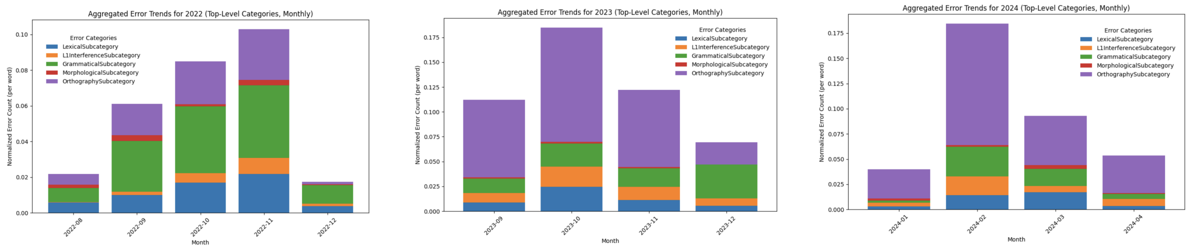
None of the observed fluctuations (e.g., rising error counts in certain months, subsequent declines) reach statistical significance (see Appendix C). However, the fine-grained L1-based labels reveal that certain patterns persist—such as Arabic speakers' difficulties with vowel representation or literal 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 the case that students do become better L2 speakers, but their assignments also become harder, leading to more errors. Alternatively, the first assignment may have been intentionally designed to be easier, resulting in fewer errors. If we exclude this initial task, we may actually observe a decline in error frequencies for the 2023 and 2024 cohorts, aligning with our original hypothesis. We plan to explore these explanations more deeply in future work, engaging with the instructors of the class as well as with the students themselves.

5.2 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



(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.

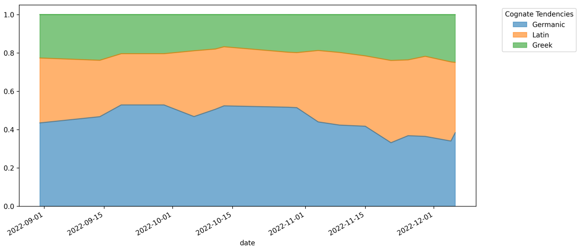


Figure 3: Proportion of Germanic, Latin, and Greek-derived vocabulary in learner writing over time (2022 cohort). The increase in Latin-based words suggests a shift toward academic vocabulary, while Germanic words remain dominant.

certain conditions, learners do shift more towards 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.

5.3 Further Syntactic Pattern Analysis

Syntactic analysis in NLP and 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 across different L1s. For example, compound noun constructions feature more prominently in Vietnamese L1 texts and much less in Chinese ones, even though one might expect the opposite due to the extensive compounding in Chinese.

We should note that the large space of possible catenae combinations and our rather sparse corpus limit 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 leave such a larger scale analysis encompassing more corpora for future work.

6 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 7. We report results in terms of accuracy and macro F1 score following the literature in this task (Goswami et al., 2024).

Models 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. We also fine-tune multiple BERT-based models on LENS namely BERT, mBERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). For these, we use a learning rate of $1e-5$ for all models and early stopping on our development set. Last, we benchmark three LLMs on LENS, 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 the fine-tuned models achieve the highest performance on LENS. 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.

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-tuining			
	GPT 4o	0.97	0.96
	LLaMa3.1	0.87	0.84
	FLAN T5	0.66	0.53

Table 7: Results of different models on the LENS dataset. LLMS require fine-tuning to outperform BERT-based and simple statistical approaches.

7 Conclusion and Future Work

We presented the first linguistically-informed LLM-based study of features of L1-L2 transfers on longitudinal data. We further presented a series of NLI experiments evaluating the model of LLMs and traditional classifiers. To achieve these goals, we complied LENS, a first-of-its-kind corpus of learner English. LENS stands apart from other similar corpora due to encompassing longitudinal data and fine-grained L1 interference annotations. We make LENS feeling available to the research community.

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

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. This perhaps limits the generalizability of our approach, but we believe that this limitation is mitigated by the fact that most second language learners opt to learn high-resource languages.

Additionally, while we conduct 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.

Ethical Considerations

The dataset collected for this research had an Institutional Review Board (IRB) review application filed at the authors' university. As the research does not involve human participants *per se*, we expect the application to qualify for exemption. The dataset was collected from writing assignments submitted by non-native English speakers enrolled in an introductory academic writing course at the authors' university. Permission for dataset use and sharing for non-commercial research purposes was obtained from the appropriate university departments. These writing samples, given that they are responses to exercise prompts, are generally non sensitive. However, careful anonymization steps have been taken to ensure that students cannot be re-identified.

During the anonymization process, meta-data that could be used to identify the student participants was eliminated. We retained only meta-data that was deemed non-identifiable: the student ages, genders, and L1s. These demographics were deemed non-identifiable on the basis that course numbers and course years are not retained, so re-identification of student participants should not be possible. The texts were then passed through an English Named Entity Recognition module[7], and all tokens that were tagged as organizations, locations, and people were replaced with a placeholder token, following Megyesi et al. (2018). Finally, the dataset was manually checked to ensure that all identifying information was removed. The final dataset does not include identifying information.

The anonymized dataset will be made available to researchers for non-commercial purposes, under a Creative Commons BY-NC-SA 4.0 license upon publication via a GitHub repository. Access will require agreement to terms that prohibit attempts to re-identify individuals or use the data for purposes beyond research.

[7]Stanford NLP, OntoNotes Release 5.0

Acknowledgments

The authors are grateful to the anonymous reviewers whose suggestions substantially improved this paper. Furthermore, the authors would like to thank Kathleen Rossell and Christina Brady at INTO-Mason for the data provided for this study. This work was partially supported by the US National Science Foundation under CAREER award 2439202.

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*.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

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.

Centre for English Corpus Linguistics (CECL). 2024. Learner corpora around the world. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. Accessed: 2025-09-17.

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.

Pieter de Haan. 2000. Tagging non-native english with the tosca-icle tagger. *Corpus Linguistics and Linguistic Theory*, pages 69–79.

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.

Susan M Gass, Jennifer Behney, and Luke Plonsky. 2020. *Second language acquisition: An introductory course*. Routledge.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of BEA*.

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*.

Dhiman Goswami, Marcos Zampieri, Kai North, Shervin Malmasi, and Antonios Anastasopoulos. 2025. Multilingual native language identification with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 193–199.

Cyril Goutte and Serge Léger. 2017. Exploring optimal voting in native language identification. In *Proceedings of BEA*.

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.

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*.

Anke Lüdeling and Hagen Hirschmann. 2015. *Error annotation systems*, page 135–158. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

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*.

Shervin Malmasi et al. 2016. *Native language identification: explorations and applications*. Ph.D. thesis, Macquarie University.

Beata Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Daniel Rosén, Carl-Johan Schenström, and Elena Volodina. 2018. Learner corpus anonymization in the age of gdpr: Insights from the creation of a learner corpus of swedish. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*, pages 47–56.

Fanny Meunier. 2016. Introduction to the longdale project. In Erik Castello, Katherine Ackerley, and Francesca Coccetta, editors, *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*, pages 123–126. Peter Lang, Berlin.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

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.

Lourdes Ortega. 2014. *Understanding second language acquisition*. Routledge.

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

Erin Pacquetet. 2024. *The Effect of Linguistic Properties on Typing Behaviors and Production Processes*. Ph.D. thesis, ProQuest LLC. ERIC Number: ED657074.

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.

Bandar Mohammad Saeed Al-Sobhi. 2019. The nitty-gritty of language learners' errors – contrastive analysis, error analysis and interlanguage. *International Journal of Education and Literacy Studies*, 7(3):49.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, and John P. McCrae. 2024. Evaluating the generalisation of an artificial learner. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 199–208, Rennes, France. LiU Electronic Press.

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.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages

147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

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.

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P. Dinu. 2017. Native language identification on text and speech. In *Proceedings of BEA*.

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 pre-trained 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 (including typos), lexical, morphological, or grammatical.

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 (including typos), lexical, morphological, or grammatical.
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)
       * *faciliter* → *facilitate* (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:

```
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 in the GitHub repository.

B Error Annotation Analysis

Table 8: Proportional Breakdown of All Error Subcategories by Assignment Length.
Legend: Orthography , L1 Interference , Lexical , Morphological , Grammatical .

Error Type	Long	Short
Orthography: Capitalization/Punctuation	0.0940	0.2134
Orthography: Consonant Doubling	0.0084	0.0310
Orthography: Consonant Substitution	0.0528	0.0534
Orthography: Hyphenation/Spacing	0.0382	0.0484
Orthography: Phonetic	0.0734	0.0720
Orthography: Punctuation	0.0015	0.0000
Orthography: Silent Letters/Irregular	0.0405	0.0658
Orthography: Typo	0.0222	0.0273
Orthography: Vowel Substitution/Omission	0.0849	0.1439
L1 Interference: Grammatical Interference	0.0245	0.0323
L1 Interference: Lexical Interference	0.0138	0.0062
L1 Interference: Orthographic Interference	0.0222	0.0571
L1 Interference: Syntactic Interference	0.0206	0.0062
Lexical: Homophone Confusion	0.0122	0.0199
Lexical: Lexical	0.1361	0.0596
Lexical: Phonological Confusion	0.0214	0.0149
Grammatical: Grammatical	0.3127	0.1315
Morphological: Morphemic Affix	0.0107	0.0099
Morphological: Overgeneralization	0.0099	0.0074

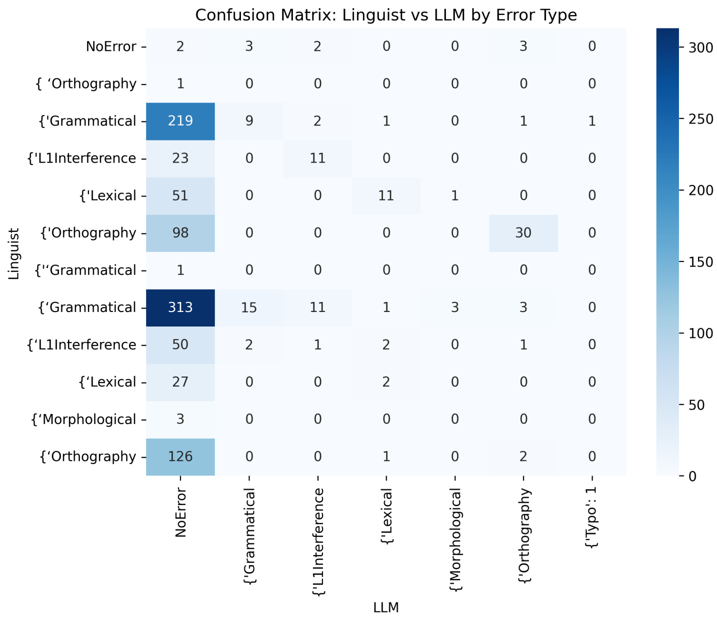


Figure 4: Confusion matrix of error type agreement between linguist and LLM.

Error Type	Capitalization/ Punctuation	Consonant Doubling	Consonant Subst.	Grammatical	Hyphenation/ Spacing	L1 Grammatical	L1 Lexical	L1 Orthographic	L1 Syntactic	Lexical	Morphological	Phonetic	Punctuation	Silent/ Irregular	Vowel Subst./ Omission	Typo
Capitalization/ Punctuation	–	(-5.05, 0.00)	(-9.76, 0.00)	(-10.48, 0.00)	(-9.36, 0.00)	(-6.96, 0.00)	(-3.60, 0.00)	(-3.89, 0.00)	(-8.10, 0.00)	(-14.62, 0.00)	(-2.80, 0.01)	(-8.09, 0.00)	(2.66, 0.01)	(-8.63, 0.00)	(-11.84, 0.00)	(-2.49, 0.02)
Consonant Doubling	(5.05, 0.00)	–	(-1.34, 0.18)	(-3.97, 0.00)	(-2.44, 0.02)	(-4.23, 0.00)	(-1.22, 0.23)	(-1.46, 0.15)	(-6.60, 0.00)	(-5.18, 0.00)	(2.78, 0.01)	(-0.94, 0.35)	(6.25, 0.00)	(-2.02, 0.05)	(-1.98, 0.05)	(0.67, 0.51)
Consonant Subst.	(9.76, 0.00)	(1.34, 0.18)	–	(-3.43, 0.00)	(-1.53, 0.13)	(-3.78, 0.00)	(-0.62, 0.54)	(-0.87, 0.39)	(-6.33, 0.00)	(-5.04, 0.00)	(5.48, 0.00)	(0.45, 0.66)	(12.61, 0.00)	(-1.02, 0.31)	(-0.82, 0.41)	(1.66, 0.10)
Grammatical	(10.48, 0.00)	(3.97, 0.00)	(3.43, 0.00)	–	(1.84, 0.07)	(-1.95, 0.05)	(1.00, 0.33)	(0.77, 0.45)	(-5.26, 0.00)	(-0.68, 0.50)	(7.56, 0.00)	(3.61, 0.00)	(12.03, 0.00)	(2.21, 0.03)	(2.95, 0.00)	(3.63, 0.00)
Hyphenation/ Spacing	(9.36, 0.00)	(2.44, 0.02)	(1.53, 0.13)	(-1.84, 0.07)	–	(-2.99, 0.00)	(0.05, 0.96)	(-0.19, 0.85)	(-5.88, 0.00)	(-2.88, 0.00)	(6.13, 0.00)	(1.82, 0.07)	(11.19, 0.00)	(0.43, 0.67)	(0.93, 0.35)	(2.46, 0.02)
L1 Grammatical	(6.96, 0.00)	(4.23, 0.00)	(3.78, 0.00)	(1.95, 0.05)	(2.99, 0.00)	–	(2.26, 0.03)	(2.08, 0.04)	(-3.71, 0.00)	(1.67, 0.10)	(5.89, 0.00)	(3.91, 0.00)	(7.45, 0.00)	(3.20, 0.00)	(3.52, 0.00)	(4.22, 0.00)
L1 Lexical	(3.60, 0.00)	(1.22, 0.23)	(0.62, 0.54)	(-1.00, 0.33)	(-0.05, 0.96)	(-2.26, 0.03)	–	(-0.18, 0.86)	(-5.24, 0.00)	(-1.38, 0.18)	(2.66, 0.01)	(0.79, 0.44)	(4.04, 0.00)	(0.16, 0.87)	(0.34, 0.74)	(1.54, 0.13)
L1 Orthographic	(3.89, 0.00)	(1.46, 0.15)	(0.87, 0.39)	(-0.77, 0.45)	(0.19, 0.85)	(-2.08, 0.04)	(0.18, 0.86)	–	(-5.12, 0.00)	(-1.14, 0.26)	(2.93, 0.01)	(1.04, 0.30)	(4.33, 0.00)	(0.40, 0.69)	(0.59, 0.56)	(1.76, 0.08)
L1 Syntactic	(8.10, 0.00)	(6.60, 0.00)	(6.33, 0.00)	(5.26, 0.00)	(5.88, 0.00)	(3.71, 0.00)	(5.24, 0.00)	(5.12, 0.00)	–	(5.13, 0.00)	(7.53, 0.00)	(6.41, 0.00)	(8.36, 0.00)	(6.00, 0.00)	(6.19, 0.00)	(6.61, 0.00)
Lexical	(14.62, 0.00)	(5.18, 0.00)	(5.04, 0.00)	(0.68, 0.50)	(2.88, 0.00)	(-1.67, 0.10)	(1.38, 0.18)	(1.14, 0.26)	(-5.13, 0.00)	–	(10.02, 0.00)	(5.10, 0.00)	(17.42, 0.00)	(3.31, 0.00)	(4.57, 0.00)	(4.36, 0.00)
Morphological	(2.80, 0.01)	(-2.78, 0.01)	(-5.48, 0.00)	(-7.56, 0.00)	(-6.13, 0.00)	(-5.89, 0.00)	(-2.66, 0.01)	(-2.93, 0.01)	(-7.53, 0.00)	(-10.02, 0.00)	–	(-4.61, 0.00)	(4.60, 0.00)	(-5.56, 0.00)	(-6.63, 0.00)	(-1.18, 0.24)
Phonetic	(8.09, 0.00)	(0.94, 0.35)	(-0.45, 0.66)	(-3.61, 0.00)	(-1.82, 0.07)	(-3.91, 0.00)	(-0.79, 0.44)	(-1.04, 0.30)	(-6.41, 0.00)	(-5.10, 0.00)	(4.61, 0.00)	–	(10.20, 0.00)	(-1.34, 0.18)	(-1.21, 0.23)	(1.38, 0.17)
Punctuation	(-2.66, 0.01)	(-6.25, 0.00)	(-12.61, 0.00)	(-12.03, 0.00)	(-11.19, 0.00)	(-7.45, 0.00)	(-4.04, 0.00)	(-4.33, 0.00)	(-8.36, 0.00)	(-17.42, 0.00)	(-4.60, 0.00)	(-10.20, 0.00)	–	(-10.35, 0.00)	(-15.69, 0.00)	(-3.11, 0.00)
Silent/ Irregular	(8.63, 0.00)	(2.02, 0.05)	(1.02, 0.31)	(-2.21, 0.03)	(-0.43, 0.67)	(-3.20, 0.00)	(-0.16, 0.87)	(-0.40, 0.69)	(-6.00, 0.00)	(-3.31, 0.00)	(5.56, 0.00)	(1.34, 0.18)	(10.35, 0.00)	–	(0.41, 0.69)	(2.16, 0.03)
Vowel Subst./ Omission	(11.84, 0.00)	(1.98, 0.05)	(0.82, 0.41)	(-2.95, 0.00)	(-0.93, 0.35)	(-3.52, 0.00)	(-0.34, 0.74)	(-0.59, 0.56)	(-6.19, 0.00)	(-4.57, 0.00)	(6.63, 0.00)	(1.21, 0.23)	(15.69, 0.00)	(-0.41, 0.69)	–	(2.08, 0.04)
Typo	**(2.49, 0.02)**	(-0.67, 0.51)	(-1.66, 0.10)	**(-3.63, 0.00)**	**(-2.46, 0.02)**	**(-4.22, 0.00)**	(-1.54, 0.13)	(-1.76, 0.08)	**(-6.61, 0.00)**	**(-4.36, 0.00)**	(1.18, 0.24)	(-1.38, 0.17)	**(3.11, 0.00)**	**(-2.16, 0.03)**	**(2.08, 0.04)**	–

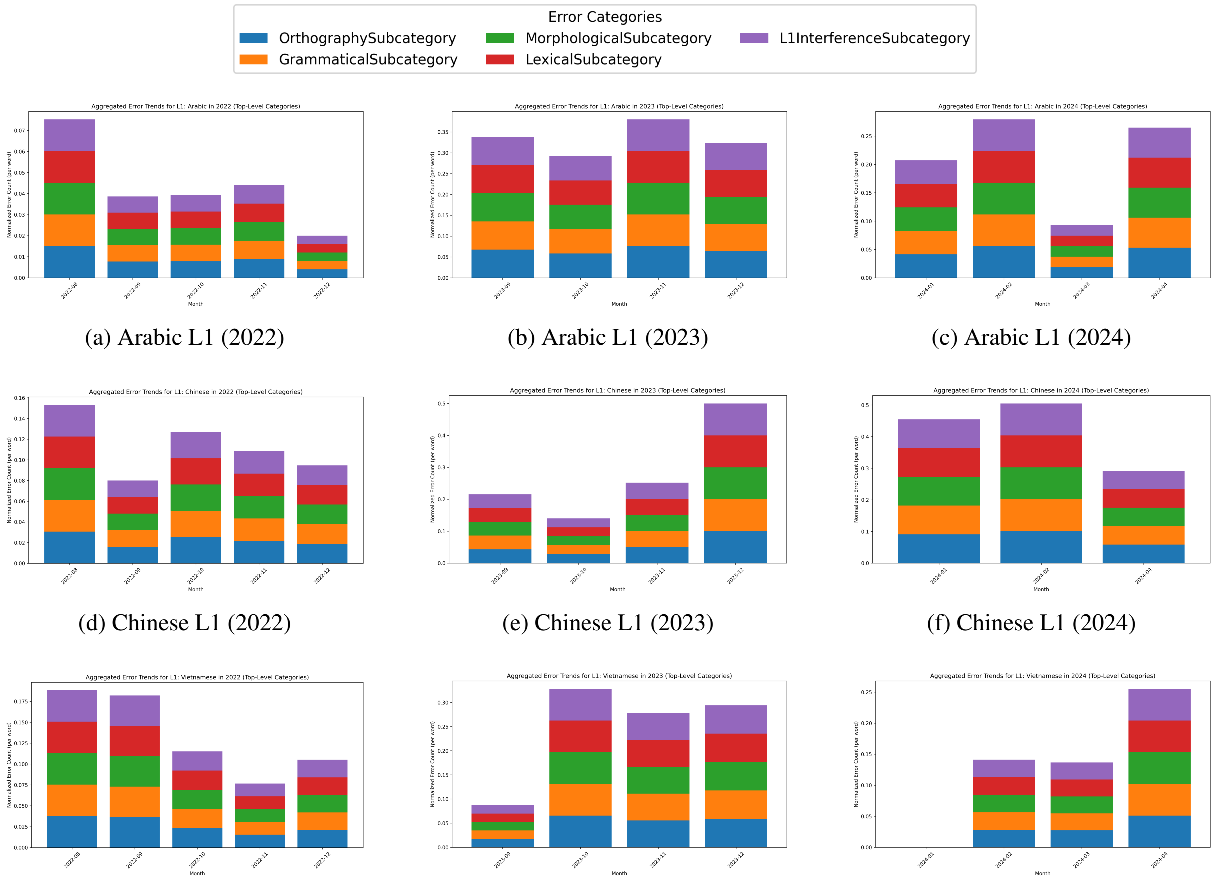
Table 9: Pairwise T-tests between error types. Each cell shows the (T-statistic, P-value). The diagonal is shown as "–".

Error Category	Precision	Recall	F1 Score
Orthography: Vowel Substitution/Omission	0.333	1.000	0.500
Grammatical: Grammatical	0.483	1.000	0.651
L1 Interference: Grammatical Interference	0.882	1.000	0.938
Orthography: Consonant Substitution	0.600	1.000	0.750
Orthography: Phonetic	0.333	1.000	0.500
Orthography: Typo	1.000	1.000	1.000
Orthography: Capitalization/Punctuation	0.778	1.000	0.875
Orthography: Hyphenation/Spacing	0.750	1.000	0.857
L1 Interference: Orthographic Interference	0.000	0.000	0.000
Lexical: Lexical	0.833	1.000	0.909
L1 Interference: Lexical Interference	1.000	1.000	1.000
Orthography: Silent Letters/Irregular	0.400	1.000	0.571
L1 Interference: Syntactic Interference	0.800	1.000	0.889
Lexical: Phonological Confusion	1.000	1.000	1.000
Morphological: Overgeneralization	0.000	0.000	0.000
Morphological: Morphemic/Affix	0.000	0.000	0.000

Table 10: Type-wise performance metrics for LLM annotations compared to human annotations.

C 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) Arabic L1 (2022) (b) Arabic L1 (2023) (c) Arabic L1 (2024)

(d) Chinese L1 (2022) (e) Chinese L1 (2023) (f) Chinese L1 (2024)

(g) Vietnamese L1 (2022) (h) Vietnamese L1 (2023) (i) Vietnamese L1 (2024)

Figure 5: Aggregated error trends by L1 and year. Each subfigure represents a different L1-year combination.

D Corpus Composition

L1	Learners	Docs	Tokens	Med. tok/doc	Entries per learner	Span(wks)	Count		Proportion	
							Long	Short	Long	Short
Arabic	35	345	63090	80.0	9.83	10.00	121	224	0.35	0.65
Azerbaijani	2	12	1934	113.0	6.00	5.00	5	7	0.42	0.58
Bengali	1	3	1990	415.0	3.00	4.00	3	0	1.00	0.00
Chinese	18	133	28678	86.5	7.33	4.00	58	75	0.43	0.57
Dari	2	26	12118	332.0	13.00	9.86	26	0	1.00	0.00
French	1	19	9139	393.0	17.00	13.86	19	0	1.00	0.00
Indonesian	1	8	820	89.0	8.00	11.00	1	7	0.13	0.87
Korean	1	14	1933	140.0	13.00	9.00	8	6	0.62	0.38
Kyrgyz	1	3	339	83.0	3.00	2.00	1	2	0.33	0.67
Portuguese	1	3	1301	281.0	3.00	4.00	3	0	1.00	0.00
Russian	1	19	12024	490.0	19.00	13.86	19	0	1.00	0.00
Sindhi	1	17	9611	567.0	13.00	13.86	15	2	0.87	0.13
Telugu	2	36	19493	416.5	18.00	13.86	35	1	0.97	0.03
Urdu	1	2	384	192.0	2.00	0.29	2	0	1.00	0.00
Vietnamese	4	47	12471	199	11.75	11.9	35	12	0.70	0.30
Total	72	687	175325	-	-	-	-	-	-	-

Table 11: Corpus composition and per-L1 breakdown, including the total number of documents, tokens, learners, and document types analyzed in this paper. Long and short categorization of the documents for this table was based on the median document length, which differs from the median document length of the subset.

Cohort Year	Num. Assignments	Num. Students	Avg. Missing
2022	19	28	10.79
2023	12	22	4.05
2024	13	38	8.26

Table 12: Longitudinal properties of the corpus, showing student engagement and data points per cohort. "Avg. Missing" refers to the average number of missing assignments per student. All subjects in a given cohort did the same writing assignment.