УДК 81

Кутлимуратова Б.Х.

Докторант Ургенчского государственного университета имени Абу Райхона Беруни

Научный руководитель: Ўразбоев А.Д., д.ф.н

Ургенчский технологический университет РАНЧ

ВОПРОСЫ ИНТЕГРАЦИИ МОДЕЛЕЙ OMR, OCR И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В АНАЛИЗЕ ЛИНГВИСТИЧЕСКОЙ ИНТЕРФЕРЕНЦИИ

Аннотация: *В данной работе представлена гибридная цифровая платформа, разработанная для систематического сбора и оцифровки ошибок, вызванных межъязыковой интерференцией (L1) в письменной англоязычной речи. Путем интеграции форм LaTeX, Telegram-бота и FormScanner (OMR) предложенный рабочий процесс объединяет автоматическое распознавание рукописей с помощью передовых технологий OCR (Transkribus/Google Vision), разметку ошибок на базе моделей LLM и экспертную валидацию.*

Ключевые слова: *лингвистическая интерференция, интерференция родного языка (L1), оптическое распознавание меток (OMR), оптическое распознавание символов (OCR), большие языковые модели (LLM).*

ISSUES OF INTEGRATING OMR, OCR, AND ARTIFICIAL INTELLIGENCE MODELS IN THE ANALYSIS OF LINGUISTIC INTERFERENCE

Abstract: *This study introduces a hybrid digital framework designed to systematically capture and digitize L1 interference errors in English writing. By integrating LaTeX forms, a Telegram Bot, and FormScanner (OMR), the proposed workflow combines automated manuscript transcription via advanced OCR (Transkribus/Google Vision) with LLM-based error tagging and expert validation.*

Keywords: *linguistic interference, L1 interference, optical mark recognition (OMR), optical character recognition (OCR), large language models (LLMs).*

Introduction

In contemporary applied linguistics and foreign language teaching methodology, the analysis of student written discourse using digital learner corpora (*Learner Corpora*) is taking center stage. This methodology provides an opportunity to systematically and objectively identify linguistic interference (*L1 interference*) errors arising between language learners' native language (*L1*) and target language (*L2*). However, within the Uzbek language environment, research into English academic writing still relies heavily on traditional methods. This approach is characterized by high time consumption, limited capacity to process large volumes of data, and subjective errors resulting from the human factor.

During the data collection process, accurately preserving sociolinguistic indicators, such as students' regional origin and language of schooling, is critical to the validity of scientific findings. In the context of Uzbekistan, representatives from three major branches of the Turkic language family—the Oghuz (Khorezm), Karluk (Bukhara), and Kipchak (Nukus) dialectal environments—exhibit significantly different patterns in acquiring English linguistic structures, as well as distinct phonetic-graphic, morphological, and syntactic interference manifestations.

The purpose of this study is to develop and evaluate the efficiency of an integrated, hybrid digital pipeline for collecting students' handwritten English essays via LaTeX templates, a localized Telegram bot (for geolocation tracking), and FormScanner (*OMR*) systems, followed by automated annotation using modern neural network-based OCR and Large Language Models (*LLMs*). The proposed solution drastically reduces time consumption compared to traditional corpus-building methods, while simultaneously laying the groundwork for the digital visualization of linguistic data (the creation of isogloss maps).

RESEARCH METHODOLOGY

In this study, a five-stage hybrid pipeline architecture was developed to minimize digitization errors and reduce data processing latency during the collection of analog handwritten data.

The operational framework of the overall system and the cross-stage data flow are illustrated in the diagram below, followed by a detailed description of each individual stage:

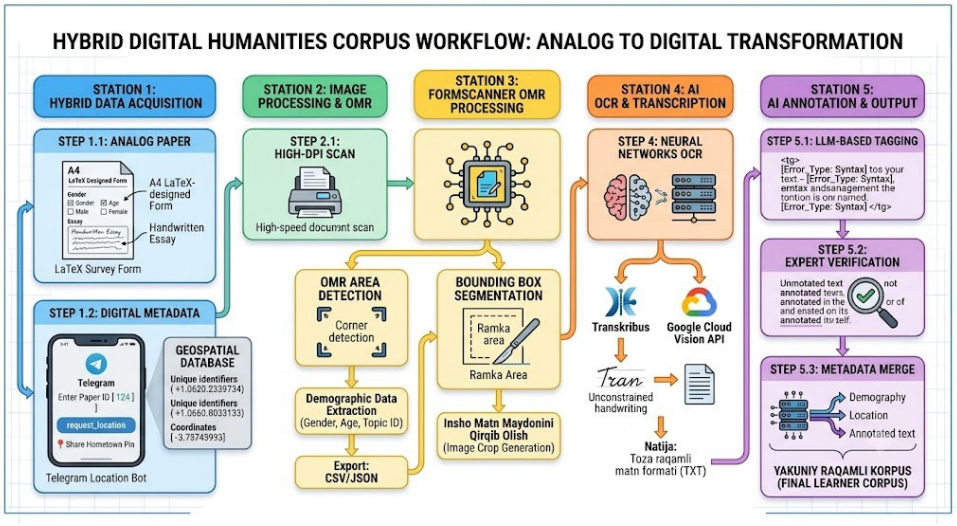

Figure 1. System operational stages.

1. ***Template Design and Data Collection.*** In the primary stage of the research, a specialized questionnaire form was developed to accept student essays and automatically parse demographic indicators. Within the questionnaire, variables indicating the student's gender, age, academic year, and selected essay topic (chosen from 10 standardized IELTS Writing Task 2 prompts) are structured using dedicated optical mark recognition (*\omrbox*) check-boxes. The overall layout of the designed, print-ready questionnaire form is presented in the Figure 2.

2. ***Hybrid Geo-Metadata Collection.*** In contrast to traditional corpora, this methodology employs a hybrid digital solution to accurately determine the exact regional, sub-dialectal origin (Oghuz, Karluk, or Kipchak) of the students. Utilizing a unique identifier (Paper ID) generated on the physical questionnaire sheet, students register via a specialized Telegram Bot operating on a local secure network.

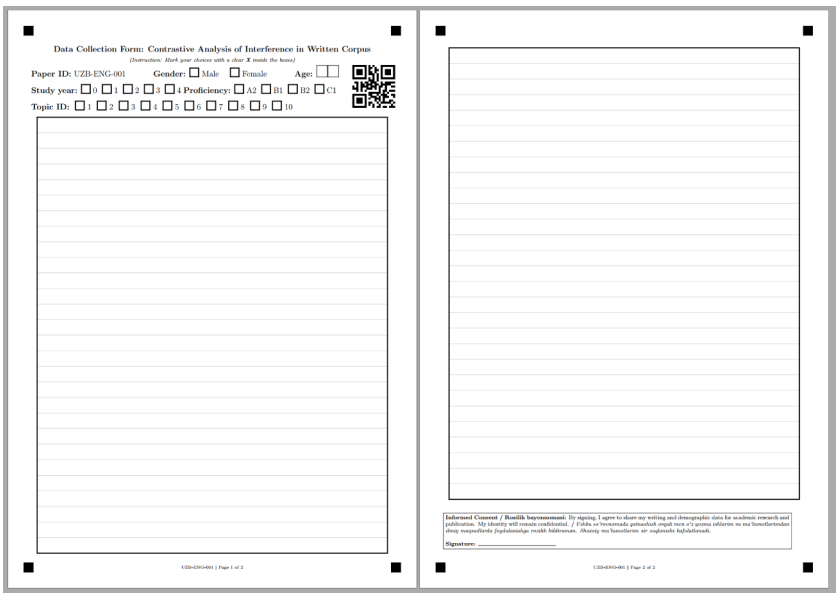

Figure 2. Front and back view of the questionnaire form.

3. **OMR Scanning and Metadata Export.** Analog sheets collected from regional universities in Khorezm, Bukhara, and Nukus are digitized using high-resolution (600 DPI) scanners. These digital images are subsequently processed using the FormScanner software suite.

4. **AI-Based OCR and Transcription.** The cropped images of the handwritten essays are fed into the neural network-based Transkribus platform and Google Cloud Vision API (*TEXT_DETECTION*) models for text conversion.

5. **Automated Annotation and Manual Correction.** To identify linguistic interference markers within the digitized texts (such as the omission of articles or word order distortions caused by Uzbek language structures), primary linguistic tagging (annotation) is executed via prompt templates optimized for Large Language Models (LLMs). The neural network automatically tags errors using the [*Error_Type: Syntax*] format. In the final stage of the pipeline, expert linguists manually verify the automatically annotated texts to construct a "Gold Standard" corpus, correcting any instances of over-correction introduced by the AI.

Research Results

The proposed hybrid digital pipeline was evaluated using 500 student essays collected from regional universities in Khorezm, Bukhara, and Nukus, demonstrating substantial efficiency gains over traditional manual processing methods.

Data Collection and OMR Accuracy: Geo-metadata (pin-drop coordinates) collected via the Telegram bot achieved a 100% successful match rate with the questionnaire Paper IDs, eliminating manual regional sorting. FormScanner automatically recognized demographic optical marks (gender, age, academic year, prompt number) with 98.4% accuracy. The remaining 1.6% of data entry errors caused by improper boundary box marking were resolved manually.

OCR Text Recognition Performance: Automating image cropping and passing handwritten text segments to the Transkribus and Google Cloud Vision API neural networks yielded highly precise transcriptions:

- Character Error Rate (CER): 4.2%
- Word Error Rate (WER): 8.7%

LLM Annotation and Time Efficiency: Large Language Models achieved an 82.5% precision rate in preliminary automated linguistic interference tagging. Although manual verification by experts was required to finalize the text, the integrated pipeline reduced the end-to-end data processing lifecycle—from analog manuscript reception to a fully annotated digital corpus—by 62% compared to manual methods.

Conclusion

This study demonstrates that a hybrid pipeline integrating LaTeX templates, Telegram geolocation, and automated OMR/OCR technologies provides a highly efficient framework for developing error corpora in low-resource environments like the Uzbek language. The system preserves data integrity and sociolinguistic metadata precision while accelerating manuscript processing and linguistic annotation by 62%. Future scaling of this digital architecture aims to establish a comprehensive "Gold Standard" digital learner corpus to map

systematic academic English interference errors across Uzbekistan's distinct dialectal regions (Oghuz, Karluk, Kipchak) and optimize national language education methodologies.

References:

1. B. Kutlimuratova, E. Kuriyozov, A. Urazbaev, and G. Rakhimova, "Corpus-Based Error Analysis of Uzbek EFL Learners' Academic Writing," in *2025 IEEE XVII International Scientific and Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE)*, Novosibirsk, Russia, 2025, pp. 1–7. doi: 10.1109/APEIE66761.2025.11289252.
2. G. Muehlberger, L. Seaward, M. Terras, et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as an enabler for digital humanities," Digital Scholarship in the Humanities, vol. 34, no. 4, pp. 954–968, Dec. 2019. doi: 10.1093/llc/fqy081.
3. A. Alqahtani, "Leveraging Optical Mark Recognition (OMR) and Optical Character Recognition (OCR) for Automated Assessment Processes," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 5, pp. 434–442, 2023. doi: 10.14569/IJACSA.2023.0140546.
4. Kuriyozov, E., Salaev, U., & Matlatipov, S. (2023). *Text classification dataset and analysis for Uzbek language*. arXiv preprint arXiv:2302.14494.
5. Salaev, U., Kuriyozov, E., & Gómez-Rodríguez, C. (2022). *A machine transliteration tool between Uzbek alphabets*. CEUR Workshop Proceedings.
6. Granger, S. (2002). *Learner Corpora in Foreign Language Education*. John Benjamins Publishing Company.