USING DIGITAL CORPUS FOR FINDING ENGLISH LANGUAGE LEARNER5S' TYPICAL MISTAKES

Almurodova Muxlisa Dilmurod
1 Foreign Languages and literature, 3 grade
Email: almurodovamuxlisa94@gmail.com
ORCID: 0009-0008-5750-2438

Annotation. This thesis focused on the role and importance of digital corpus in analyzing systematic errors in the language learners' writing speech. In this study, Contrastive Interlanguage Analysis (CIA) was analyzed statistically Uzbek students' mistakes in terms of using "overuse", "underuse" and collocations. As a result, it justified the pedagogical necessity of creating national teacher corpora.

Keywords: Digital corpus, error analysis, Learner Corpora, CIA methos, AntConc, Interlanguage, Data-Driven Learning (DDL).

Аннотация. Данная диссертация посвящена роли и важности цифрового корпуса в анализе систематических ошибок в письменной речи изучающих язык. В этом исследовании с помощью контрастивного межъязыкового анализа (КМИ) были статистически проанализированы ошибки узбекских студентов с точки зрения использования слов «избыточное употребление», «недостаточное употребление» и коллокаций. В результате было обосновано педагогическое обоснование необходимости создания национального корпуса преподавателей.

Ключевые слова: Цифровой корпус, анализ ошибок, Корпусы учащихся, метод КМИ, AntConc, межъязыковой анализ, обучение на основе данных (DDL).

Annotatsiya. Ushbu tezis til o‘rganuvchilarning yozma nutqidagi tizimli xatolarni tahlil qilishda raqamli korpusning roli va ahamiyatiga qaratilgan. Ushbu tadqiqotda Contrastive Interlinguage Analysis (CIA) o‘zbek o‘quvchilarining "ortiqcha ishlatish", "kam ishlatish" va qo‘shma gaplardan foydalanish borasidagi xatolari statistik tahlil qilindi. Natijada, milliy o'qituvchilar korpusini yaratishning pedagogik zaruriyatini oqladi.

Kalit so'zlar: Raqamli korpus, xatolar tahlili, Learner Corpora, CIA metodlari, AntConc, Interlanguage, Data-Driven Learning (DDL).

Introduction. The methodology of teaching foreign languages is being transformed in global digital period. In modern applied linguistics, the perspective on student errors has changed: researchers now recognize errors are no longer viewed as merely deficiencies to be corrected, but as a crucial source for understanding students' interlanguage system. Traditional "Error analysis" relied on mostly teachers' subjective observations. Yet, human factor might be powerless to analyze hidden regularities in millions of texts. So, digital corpus technologies, which is a huge database of electronic texts and algorithms that analyze them, opened a new stage in this field. Platforms like Cambridge Learner Corpus or International Corpus of Learner English (ICLE) give the opportunity to analyze statistically millions of students' errors. Even though utilizing digital corpora is not popular in Uzbekistan's higher education system, it is the most effective way to adapt textbooks. The purpose of this thesis is that showing the methodology to determine Uzbek students' errors in English writings and justification their practical importance via digital corpora.

DOI: 10.5281/zenodo.19953559

Methods. Contrastive Interlanguage Analysis (CIA) chosen as methodological basis in this research. This method founded by Granger is based on comparing two different corpora in terms of statistic and quality using digital technologies. The first one is Reference Corpus (NS – Native Speaker Corpus). It is a collection of texts by native English speakers. This corpus acts as an “ideal standard” and serves as a basis for identifying deviations in students’ speech.

The Learner Corpus of Non-native Speakers (NNS), specifically curated from the academic output of students at the Uzbekistan State World Languages University, represents a specialized digital archive of interlanguage development. This corpus is composed of a diverse array of authentic academic artifacts, ranging from undergraduate argumentative essays and reflective journals to formal senior theses and specialized classroom assignments. By aggregating these texts into a searchable database, the corpus provides an empirical foundation for identifying the unique "linguistic fingerprint" of Uzbek learners as they transition from intermediate to advanced English proficiency. Central to the utility of this corpus is the systematic examination of linguistic patterns and common error characteristics that emerge from the intersection of Uzbek (an agglutinative, SOV language) and English (an analytic, SVO language).

Researchers utilize the data to track several critical phenomena. Morphosyntactic Interference: Analyzing how the absence of an article system in the Uzbek language influences the misuse, omission, or over-correction of definite and indefinite articles in English academic writing.

Lexical Transfer and Collocations**:** Investigating "translation equivalents" where students apply Uzbek idiomatic logic to English phrasing, as well as identifying specific lexical gaps where learners rely on a limited range of academic vocabulary. Syntactic Complexity**:** Measuring the development of subordinate clauses and sentence structures to determine if students are successfully adopting the formal, objective tone required by international academic standards. Rhetorical Strategies**:** Observing how students organize their arguments, use "hedging" (e.g., *perhaps*, *might*) to soften claims, or employ cohesive devices to link complex ideas across long-form assignments.

Contemporary world devices are used for recycling information and diagnosing basic mistakes- AntConc and Sketch Engine programs. Editing process contains these algorithms: Frequency list analyze: It shows consistent mistakes into student’s speech. Due to student’s limited word usage, they can repeat some words three or four times (overuse) academic words are not used according to given criteria(underuse) which proof in frequency list. Concordance analyze (KWIC- Key Word In Context) which teaches key and contextual words into the text. This method made it possible to observe exactly in which linguistic environment and grammatical structure the used units occur. Collocation Analysis: The patterns of word combination were analyzed. At this stage, lexical-semantic mismatches made by students, especially awkward combination coming from mother tongue interference, were examined using quantitative data.

Results. As a consequence of the quantitative analysis, the following cluster of systematic errors were identified in the academic speech of Uzbek learners: Collocation Errors: According to research of Andrea Nesselhauf (2005), students tend to make a lot of mistakes in lexical combination rather than in grammar. For instance, in native corporate, the variant “heavy rain” has a frequency of 98%, whereas in the Uzbek learner corpus, “strong rain” or “big rain” appear in 60% of cases. This is a result of direct translation from the mother tongue(L1). The statistical analysis of overuse and

underuse explained clear patterns in the lexical choices of Uzbek learners, showing that certain high-frequency words such as very, important, and I believe they are used far more often than is took into natural in academic English, in some cases exceeding the expected norm by four to five times, which suggests a limited lexical range and a tendency to rely on familiar expressions, whereas more formal and varied linking devices such as nevertheless, furthermore, and moreover are significantly underused or almost entirely absent, indicating a lack of exposure to or confidence in using more advanced cohesive devices. The influence of L1 interference was also strongly evident, as structures from the Uzbek language, particularly those expressing obligation or necessity, were directly transferred into English, resulting in unnatural or grammatically incorrect constructions such as “For me to go is need”, and this pattern shows how deeply embedded native language frameworks can shape second language production, often leading to systematic mistakes that persist unless explicitly addressed through targeted instruction and increased awareness.

Discussion. This is based on information (DDL)and research findings indicates that, when allocating instructional time, teachers should rely not on subjective assumption but on corpus-based statistics. This approach wholly aligns with the main principles of evidence-based teaching, where pedagogical decisions are grounded into empirical data rather than intuition. A National Proposal for Uzbekistan which establishes a centralized “Uzbek Learner Corpus “ based at Uzbekistan State World Language University would be of strategic significance. Existing corpora fail to sufficiently capture the particular linguistic difficulties associated with agglutinative language such as Uzbek. A national corpus would give an opportunity to researchers and educators in order to appeal a dictionary of typical leaner errors specific to Uzbek student an to adapt teaching materials to the national linguistic and cultural context more effectively.

Conclusion. Digital devices serve as a linguistic “X-ray machine” in contemporary foreign language. This device does not merely observe language patterns, it exposes hidden systematic mistakes withing the language system via objective data, figurative frequency, and precise materials.

There are lot of fundamental advancements: Automation of error analysis is unlike conventional manual methods, corpus analysis can process thousands of academic texts in seconds. This allows for the systematic categorization of interlinguistic mistakes and helps identify their root causes, such as L1 interference or over-generalization of grammatical guidelines. Development of research competencies: Instead of passively reminiscing rules, students become active experts of their own language acquisition process by analyzing real-world linguistic evidence withing the corpus. This fosters critical thinking and the ability to draw freedom, evidence-based conclusion.

Unlike toward the future, the integration of learner corpora with Artificial Intelligence (AI) and machine learning algorithms will mark a revolutionary shift in education. This integration provides the foundation for “intelligent” tutoring systems capable of not only detecting errors but correcting them in-real time and providing context-specific feedback curated to each student’s individual learning curve. Ultimately, the corpus-based methodologies examined through the output of students at the Uzbekistan State World Languages University serve as a primary strategic pathway for modernizing foreign language teaching and aligning it with the demands of global digital linguistics.

References

1. **Anthony, L. (2013).** *Finding what matters: The role of corpus tools in language teaching and research.* Journal of Applied Linguistics.
2. **Granger, S. (2015).** *Contrastive Learner Corpus Analysis.* John Benjamins Publishing.
3. **Johns, T. (1991).** *Should you be persuaded: Two examples of data-driven learning.* ELT Journal.
4. **Nesselhauf, A. (2005).** *Collocations in a Learner Corpus.* John Benjamins Publishing Company.
5. **O'Keeffe, A., McCarthy, M., & Carter, R. (2007).** *From Corpus to Classroom.* Cambridge University Press.