METHODOLOGICAL PRINCIPLES OF DEVELOPING A NATIONAL LEARNER ERROR CORPUS FOR AUTOMATED ESSAY SCORING SYSTEMS IN UZBEKISTAN

Kholbutaeva Pokiza
Master's Student of Webster University in Tashkent
pokizaxolbutayeva@gmail.com

Abstract. The development of automated essay scoring systems requires reliable linguistic resources that reflect the actual writing patterns of language learners. However, automated assessment tools trained primarily on international datasets may fail to identify language-specific errors and contextual features characteristic of learners in Uzbekistan. This thesis examines the methodological principles for developing a national learner error corpus to support automated essay scoring in Uzbek educational contexts. The proposed framework focuses on representative data collection, learner proficiency classification, systematic error annotation, linguistic diversity, ethical data management, and corpus validation. Particular attention is given to recurring grammatical, lexical, orthographic, and discourse-level errors in English essays written by Uzbek learners. The study proposes an annotation framework that combines expert linguistic judgment with computational processing. Establishing such a corpus could improve the reliability, contextual relevance, and educational usefulness of automated essay scoring systems while supporting the development of locally adapted language assessment technologies.

Keywords: learner error corpus, automated essay scoring, natural language processing, language assessment, corpus linguistics, Uzbekistan, educational technology, error annotation.

MAIN PART

The development of a national learner error corpus in Uzbekistan requires a systematic approach to data collection, linguistic annotation, and computational analysis. The corpus should include English essays written by learners from different educational institutions, regions, and proficiency levels. Each text should be analyzed according to standardized categories, including grammatical, lexical, spelling, punctuation, and syntactic errors. Particular attention should be given to linguistic difficulties that may arise from differences between Uzbek and English. However, these patterns must be identified through empirical analysis rather than assumptions. To ensure reliability, trained linguists should annotate the texts using consistent guidelines, while independent evaluation should be conducted to measure annotation agreement. The corpus should also include relevant metadata and, where possible, expert-assigned essay scores to support the development of automated essay scoring systems. Ethical data collection, anonymization, and secure storage are essential components of the methodology. Finally, the effectiveness of the corpus should be evaluated by comparing automated assessment results with human expert judgments. This approach could contribute to more accurate, transparent, and context-sensitive essay assessment in Uzbekistan.

Purpose and methodology. The purpose of this thesis is to identify the principal methodological requirements for developing a national learner error corpus that can support automated essay scoring systems in Uzbekistan. The proposed methodology combines corpus linguistics, contrastive linguistic analysis, expert annotation, and computational language processing. First, written samples should be collected from learners representing different educational institutions, regions, age groups, and English proficiency levels. Second, each essay should be accompanied by relevant metadata,

including the learner's proficiency level, writing task, educational context, and assessment conditions, while excluding unnecessary personally identifiable information. Third, errors should be classified according to a consistent annotation scheme covering spelling, punctuation, morphology, syntax, lexical choice, grammatical agreement, and discourse organization. Fourth, trained annotators should independently label a representative portion of the corpus, and inter-annotator agreement should be calculated to evaluate annotation consistency. Finally, the corpus should be divided into training, validation, and testing subsets to support the reliable development and evaluation of automated scoring models.

Results and discussion. The proposed methodological framework identifies five interconnected principles for constructing an effective national learner error corpus. The first principle is representativeness: the corpus should reflect the diversity of English learners in Uzbekistan rather than relying exclusively on samples from a single institution or proficiency group. The second is annotation consistency, which requires clear definitions of error categories, standardized correction procedures, and documented annotation guidelines. The third is linguistic sensitivity, particularly the recognition of recurring difficulties associated with the transfer of Uzbek linguistic structures into English writing. However, potential language-transfer effects should be established through empirical analysis rather than assumed in advance. The fourth principle is pedagogical relevance: annotated errors should be linked, where feasible, to proficiency levels and writing assessment criteria so that corpus information can support meaningful feedback. The fifth principle is ethical and technical reliability, including informed consent, secure data storage, transparent data governance, and independent evaluation of the resulting models. Together, these principles could provide a foundation for developing automated essay scoring tools that are better

adapted to the linguistic and educational conditions of Uzbekistan. An additional methodological consideration is the distinction between error detection and essay scoring. Error detection identifies potentially incorrect language forms, whereas essay scoring evaluates broader dimensions of writing quality, including coherence, task fulfillment, organization, and argument development. Therefore, the proposed corpus should not be limited to isolated grammatical mistakes. It should also preserve essay-level context and, where appropriate, include expert ratings based on an explicitly defined assessment rubric. Such a design would enable researchers to investigate the relationship between linguistic accuracy and overall writing proficiency without treating the frequency of errors as the sole indicator of writing quality. Model evaluation should include both technical performance measures and human-centered assessments of feedback usefulness, fairness, and interpretability.

Conclusion. Developing a national learner error corpus represents a potentially valuable step toward improving automated essay scoring in Uzbekistan. Its methodological foundation should include representative sampling, standardized annotation, empirically verified linguistic analysis, educationally meaningful metadata, and rigorous model validation. Collaboration among linguists, language teachers, assessment specialists, and computational researchers would help ensure that the corpus reflects both linguistic realities and practical educational requirements. The inclusion of essay-level information alongside detailed error annotations could support more comprehensive assessment than systems focused exclusively on grammatical accuracy. Furthermore, transparent ethical procedures and regular corpus updates would strengthen the reliability and long-term usefulness of the resource. Future empirical research should pilot the proposed framework, establish annotation reliability, and test

whether corpus-informed models provide more accurate and useful assessments than models trained on less context-specific datasets.

References

1. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3).
2. Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In *The Handbook of Automated Essay Evaluation* (pp. 55–67). Routledge.
3. Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching, 5*(4), 161–170.
4. Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480.