Corpus-Based Structural and Semantic Analysis of IT Educational Terminology in a Bilingual English-Uzbek Corpus

Malika Tilavova[1], Oybek Akhmedov[2], Handoko Hum[3] and Lobar Sultonova[4]
[1]*Jizzakh State Pedagogical University, Sharof Rashidov Str. 4, 130100 Jizzakh, Uzbekistan*
[2]*Uzbek State World Languages University, Kichik Halka Yo'li Str. 21A, 100173 Tashkent, Uzbekistan*
[3]*Universitas Andalas, Jl. Dr. Mohammad Hatta, 25175 Padang, West Sumatra, Indonesia*
[4]*Tashkent State University of Law, Sayilgoh Str. 35, 100047 Tashkent, Uzbekistan*
tilavovamalika005@gmail.com, uzdjtu@uzswlu.uz, handoko@hum.unand.ac.id, tdyu@exat.uz

Keywords: Educational Terminology, Corpus Linguistics, Natural Language Processing, Terminology Extraction, Bilingual Alignment, Information Technology Education, Neologism Detection.

Abstract: This study provides a reproducible, corpus-based framework for the structural and semantic analysis of educational terminology related to information technology in a bilingual (English-Uzbek) context. A comprehensive bilingual corpus containing approximately 1.2 million tokens (620,000 in English and 580,000 in Uzbek) from information technology textbooks and educational resources was constructed. A carefully selected and manually annotated evaluation set of 5,000 lexical items from this corpus served as the basis for quantitative evaluation of the extraction and classification models. Terminology candidates were automatically extracted from the full corpus using a hybrid statistical and linguistic algorithm combining TF-IDF, C-value/NC-value, and morphosyntactic filtering. Evaluation on a carefully selected set showed high reliability: precision = 0.91, recall = 0.87, and F1 score = 0.89. A prototype bilingual matching tool using FastText subword embeddings / vectors and cosine similarity achieved an accuracy of 0.86 for matching English and Uzbek terms. In addition, a supervised XGBoost classifier was trained on the evaluation set to detect neologisms and metaphorical extensions of computer terminology, achieving an F1 score of 0.88. The proposed system provides a repeatable natural language processing pipeline for extracting, matching, and semantic classification of bilingual terminology. It serves as both a quantitative benchmark and a practical tool for integrating computer vocabulary into educational resources. This approach facilitates the systematic study of the structural and semantic dynamics of bilingual educational lexicons, thereby supporting future research and the development of practical curricula.

1 INTRODUCTION

The rapid integration of digital technologies into educational systems has profoundly transformed pedagogical discourse and its lexical structure. As information technologies play an increasingly prominent role in teaching, learning, assessment, and academic communication, new terminological units are emerging, while existing concepts are undergoing structural and semantic transformations.

In bilingual educational contexts, particularly those using English and Uzbek, this process is amplified by interlinguistic interaction, assimilation, and morphological adaptation.

English is the primary language of technological innovation worldwide. Consequently, most pedagogical terms related to information technologies originate in English and are subsequently integrated into other linguistic systems. In the Uzbek education sector, this integration leads to the coexistence of direct assimilation, adapted lexical forms, mixed constructions, and hybrid formations. These changes reflect not only lexical expansion but also a systematic structural reorganization and semantic enrichment of bilingual pedagogical vocabulary.

Despite the growing academic interest in the development of digital discourse and terminology, existing research in the English-Uzbek context remains primarily descriptive and qualitative. Quantitative, corpus-based, and computationally reproducible approaches remain limited. In particular, no study combines structural analysis, semantic modeling, and automatic terminology

extraction within a single natural language processing (NLP) system.

Previous studies conducted in the English-Uzbek educational context have primarily used qualitative methods, lacking reproducible computer processing chains, quantitative assessments, or bilingual matching systems. To address these shortcomings, this study proposes a corpus-based natural language processing framework for analyzing the structural and semantic characteristics of bilingual educational terminology, combining the construction of a reproducible bilingual corpus and morphosyntactic tagging, a hybrid terminology extraction algorithm evaluated using accuracy, recall, and F1 score metrics, a prototype English-Uzbek bilingual terminology matching system and a supervised classification model for detecting neologisms and metaphorical semantic extensions.

The study of automatic terminology extraction has a long history. The C-value/NC-value method has proven effective in identifying multi-component terms [1], and corpus-based statistical term extraction methods have also shown high performance [2]. Semantic structure and taxonomy extraction have been systematically evaluated in SemEval tasks [3]. Natural language processing (NLP) algorithms and frameworks provide fundamental approaches to terminology modeling [4], etc. Lexico-semantic resources such as WordNet serve as a conceptual framework for modeling semantic relationships between terms [5].

This approach provides empirical insights into structural and semantic changes, as well as practical computational tools for bilingual information technology education.

2 METHODS

2.1 Conceptual Framework and Research Design

In this study, a corpus-based computational linguistics approach is combined with supervised machine learning. The research design is based on a three-stage hybrid architecture:

1) Automatic terminology extraction using a hybrid model.
2) Semantic matching of English-Uzbek term pairs (bilingual matching).
3) Classification of neologisms and metaphorical units using XGBoost.

The theoretical framework combines terminology theory, principles of structural-semantic analysis, and statistical text processing methods. The methodological model combines linguistic rules and probabilistic indicators.

2.2 Corpus and Evaluation Set

The corpus construction and evaluation methodology consisted of the following stages:

1) Corpus Structure. A bilingual training corpus of approximately 1.2 million tokens (English: 620,000; Uzbek: 580,000) was created from textbooks, machine learning resources, terminology dictionaries, and academic texts. The entire corpus served as a source for extracting candidate terms. From this corpus, a smaller, carefully selected evaluation set of 5,000 manually annotated lexical items, including 500 information technology-related terms, was created as a reference for quantitative evaluation. As a result of this difference, the full corpus and the evaluation set are not identical.
2) Preprocessing. The corpus underwent the following standard preprocessing steps: tokenization, lemmatization, word denoising, morphological segmentation, and multi-unit detection. These steps reduced noise and increased the accuracy of candidate term detection.

2.3 Hybrid Term Extraction Model

A hybrid model combining statistical and linguistic components was developed:

1) Statistical component:
   - Extraction of n-grams (bigrams, trigrams);
   - TF-IDF weight and frequency analysis;
   - C-value/NC-value method for evaluating multi-component terms.
2) Linguistic component:
   - Extraction of structural patterns (Adj+N, N+N, N+Prep+N);
   - Extraction of affixal and derivational forms;
   - Term extraction.

The model was tested on a small set of 5000 terms; the results are presented in Section IV.

2.4 Bilingual Term Matching

Matching algorithm: lexical similarity (based on distance), contextual co-occurrence frequency,

translation equivalence, and structural similarity. Term pairs were generated semi-automatically and validated by experts. A subset of 1200 bilingual term pairs was used for evaluation.

2.5 Classification of Neologisms and Metaphors

XGBoost was used to identify innovative lexical units in information technology education.

- Characteristics. Morphological structure, number of components, assimilation, semantic expansion, and contextual dispersion.
- Model optimization. 80/20 split between training and test sets, 5-fold cross-validation, and optimized hyperparameters through exhaustive search.

2.6 Statistical Evaluation Criteria

The following metrics were used to evaluate the performance:

$$\text{Precision} = TP/(TP + FP),$$
$$\text{Recall} = TP/(TP + FN),$$
$$F1 = 2 \times P \times R/(P + R).$$

All calculations were performed in Python.

2.7 Corpus Structure and Lexical Unit Selection

In addition to the 5,000-item evaluation set used for the quantitative extraction, matching, and classification metrics reported above, a separate, smaller subset of 1,000 lexical units was selected from the entire corpus (approximately 1.2 million tokens) specifically to conduct an in-depth qualitative analysis of semantic phenomena such as complex structures, affix formation and lexical acquisition, neologism formation, and metaphorical extensions. This approach ensures reproducibility and quantitative accuracy in the analysis of bilingual terminology used in the fields of education and information technology. For instance, multi-compound terms consist of several word combinations that form a single meaning. For example, "*Learning Management System*" (Ta'limni boshqarish tizimi) is a three-word expression, using structure noun+noun+noun in English, and the same sequence is retained in Uzbek. In the examples "*Cloud-Based Learning Platform"* (Bulutli ta'lim platformasi) and "*Virtual Classroom Environment* " (Virtual sinf muhiti) the modifiers define the core of the term and are used as the terminological unit. *LMS* (Learning Management System) is translated into Uzbek as *TBT* (Ta'limni boshqarish tizimi). Similarly, *MOOC* (Massive Open Online Course) and *API* (Application Programming Interface) [6]. Hybrid forms, the expressions *e-ta'lim platformasi* or *online darslik* where the prefix or beginning of the English word is combined with the Uzbek base. This form often appears under the influence of computer innovations. *EdTech, cyberclass, simulator.* These words have intrinsic meaning and form a minimal structural unit. *Digitization → digitization, virtualization → virtualization, education → educational*. Affixes play an important role in determining terminological meaning.

The term extraction module was developed using the C-value/NC-value approach and corpus-based statistics [2]. Vector images and nesting models [7] provided a theoretical basis for estimating semantic similarity. The XGBoost gradient boosting algorithm [8] was used for classification.

2.8 Reliability and Limitations

The reliability mechanisms include expert verification, cross-validation, and comparison with a reference corpus. Limitations include the relatively small size of the corpus, partial subjective semantic evaluation, and the incomplete integration of deep lexical additions.

3 RESEARCH RESULTS

In this study, the proposed hybrid architecture was evaluated on three main modules: term extraction, bilingual matching, and neologism/metaphor classification. The results show that the model performs well and is balanced in both structural and semantic processing.

3.1 Hybrid Term Extraction

A hybrid term extraction model combining TF-IDF, C/NC meaning, and morphosyntactic analysis was applied to the entire corpus (approximately 1.2 million tokens). The evaluation was performed on a subset of 5000 manually annotated terms. The detailed results are presented in Table 1.

Table 1: Performance of hybrid terminology extraction model.

Metric	Value
Precision	0.91
Recall	0.87
F1-score	0.89

According to the first table we can see these results:

- Precision (0.91). Indicates that the majority of extracted terms are relevant, demonstrating the effectiveness of the combination of statistical and linguistic rules.
- Recall (0.87). Indicates that most reference terms are correctly identified, although some rare or highly variable terms are omitted.
- F1 score (0.89). Demonstrates a good balance between precision and recall, indicating the robustness of the model for bilingual learning content. • The hybrid approach is particularly effective for multi-component terms due to the integration of C-value/NC-values and grammatical categories.

Advantages:

- Effectively identify multi-component terms that frequently occur in educational content related to information technology.
- Combination of complementary approaches. Statistical estimation identifies frequently occurring and context-dependent terms, while linguistic rules identify structurally complex forms.

Disadvantages:

- Rare, highly variable, or idiomatic terms may sometimes be omitted.

This suggests that the hybrid approach is more robust than purely statistical or rule-based extraction for bilingual educational corpora.

3.2 Bilingual Matching

The bilingual matching model uses cosine similarity in FastText subword embeddings / vectors, in conjunction with structural and contextual features, to match English and Uzbek terms. The result is shown in Table 2.

Table 2: Performance of bilingual alignment model.

Evaluation Metric	Value
Accuracy	0.86

The bilingual matching accuracy was 0.86, which is slightly lower than that of the extraction and classification modules due to the following reasons:

1) Lack of terminological equivalence. Computer terminology is mainly borrowed from English, and many units have been standardized in this language as part of an international debate. In Uzbek:
   - Some terms are introduced through direct borrowing (for example, transliteration or phonetic adaptation).
   - Others are presented as explanatory translations.
   - Some do not have a stable equivalent.

   In the absence of full equivalence, semantic compatibility is not the same. For example, an English term may cover a wide conceptual area, while its Uzbek version may have a more limited or, conversely, more general meaning.

   Therefore, even if the model detects structural or lexical similarity, it cannot guarantee complete compatibility at the conceptual level. This leads to a decrease in accuracy.
2) Partial equivalence. In cases of partial equivalence, only part of the semantic space of two terms coincides. This happens in the following cases:
   - The term is multifunctional in one language, and highly specialized in another.
   - In one language, the term is stable in professional speech, and in another language it is still in the process of terminological formation.

   In this case, the model can determine the correspondence based on structural and lexical features, but it becomes more difficult to determine semantic completeness. In particular, in matching mechanisms based on rules or dictionary surfaces, partial correspondences are often evaluated as complete, and vice versa.

Thus, the accuracy score of 0.86 is explained by the high proportion of partial correspondences.

Contextual semantic variation. Many computer terms exhibit polysemantic or contextual variations in meaning. For example:

- The term can be used only in a technical context;
- In a pedagogical context - in a broad or metaphorical sense.

However, during adaptation, the term is often evaluated separately. If the context is not sufficiently taken into account:

- The meaning is poorly generalized,
- Or an inappropriate semantic pair is chosen.

This is especially evident for neologisms and metaphors, since their conceptual structure is not yet fully stable. As a result, structural similarity (for example, morphological parallelism) does not guarantee semantic completeness.

This result shows that structural similarity does not always mean semantic completeness. Conceptual differences are observed, especially in neologisms and metaphors. In the future, integrating semantic similarity models based on placement (e.g., context vectors) may improve accuracy.

The results obtained confirm the effectiveness of statistical and linguistic integration in term extraction [1]. The results of bilingual alignment are explained by the complexity of semantic equivalence [9], [10]. The use of grafting models [11] increases the accuracy of the alignment.

3.3 XGBoost Model

The XGBoost-based neologism and metaphor classification model was assessed using precision, recall, and F1-score. The evaluation outcomes are presented in Table 3.

Table 3: Performance of xgboost-based neologism and metaphor classification.

Metric	Value
Precision	0.89
Recall	0.87
F1-score	0.88

The XGBoost model achieved an F1 score of 0.88. The high precision (0.89) indicates that the model has few false positives. The recall score (0.87) indicates that a high proportion of innovative units were detected.

The results show that:

- Morphological renewal (affixation, word formation) can be detected automatically;
- Metaphorical expansion is highly context-dependent.

Errors were mainly observed in ambiguous units, which highlights the need for a stronger semantic disambiguation mechanism.

A comparison of the results of the three modules revealed the following:

- The term extraction module showed the highest F1 score (0.89).
- The classification module showed consistent results (F1 = 0.88).
- The bilingual matching accuracy (0.86) was slightly lower than that of the other modules, which can be explained by the complexity of semantic equivalence.

Overall, the hybrid approach turned out to be effective for automatic processing of IT terminology, both in terms of structure and semantics.

According to the module results interrelation the three modules worked as an integrated system:

- High term separation accuracy provided high-quality input data for subsequent modules.
- The adaptation module established the bilingual equivalence of the terminology database.
- The classification module conceptually separated innovative units.

Therefore, the overall efficiency of the system was greater than the sum of the efficiency of each individual module.

So, the scientific and practical significance of this study is huge.

The results obtained allow us to draw the following scientific conclusions:

1) The combination of statistical and linguistic approaches is effective for terminology separation.
2) Semantic depth is a decisive factor in the adaptation of bilingual terminology.
3) Machine learning models provide stable results for identifying innovative lexical units.

In practice, the model can be used for:

- Automatic updating of computer and educational dictionaries,
- Standardization of terminology,
- Development of bilingual educational resources.

4 CONCLUSIONS

This study developed and evaluated a corpus-based natural language processing (NLP) system for structural and semantic analysis of educational terminology related to information technology in a bilingual (English-Uzbek) context. The study showed that the integration of statistical extraction methods (TF-IDF, C-value/NC-value) with linguistic filtering (morphosyntactic constraints) provides reliable

terminology identification in a large number of educational corporations. Quantitative evaluation against a manually annotated reference set confirmed the robustness of the extraction model (F1 = 0.89), indicating high terminological accuracy and consistent recall performance.

A prototype of a bilingual matching system based on FastText augmentation and cosine similarity achieved an accuracy of 0.86, confirming the feasibility of semi-automated English-Uzbek terminology matching in specialized educational fields. Furthermore, the classification guided by XGBoost effectively identified neologisms and metaphorical extensions of the term "IT terminology" (F1 = 0.88), revealing measurable patterns of semantic innovation within the corpus.

Overall, the results confirm that the integration of "informatics" into educational discourse leads to complex and multi-component lexical formations, as well as observable processes of semantic expansion. The proposed project not only provides a quantitative reference base for bilingual terminology research, but also a practical tool for curriculum development, textbook creation, and lexicographic standardization.

Despite these promising results, the project has some limitations, such as the size of the corpus and the degree of subjectivity in semantic annotation. Future studies could expand the corpus, introduce deep contextual annotations, and explore interdisciplinary terminological transfer to improve the scalability and generalizability of the results.

REFERENCES

[1] K. T. Frantzi, S. Ananiadou, and H. Mima, “Automatic recognition of multi-word terms: The C-value/NC-value method,” International Journal on Digital Libraries, vol. 3, no. 2, pp. 115-130, 2000, [Online]. Available: https://doi.org/10.1007/s007999900023.

[2] P. Drouin, “Term extraction using non-technical corpora as a point of leverage,” Terminology, vol. 9, no. 1, pp. 99-115, 2003.

[3] G. Bordea, P. Buitelaar, S. Faralli, and R. Navigli, “SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval),” in Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 2015, pp. 902-910, [Online]. Available: https://doi.org/10.18653/v1/S15-2151.

[4] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., draft. Stanford, CA, USA: Stanford University, 2023, [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[5] T. McEnery and A. Hardie, Corpus Linguistics: Method, Theory and Practice. Cambridge, U.K.: Cambridge University Press, 2012.

[6] M. M. Tilavova, “The process of word formation in modern English and the types of formation of words related to education,” Mental Enlightenment Scientific-Methodological Journal, vol. 4, no. 6, pp. 302-313, 2023, [Online]. Available: https://doi.org/10.37547/mesmj-V4-I6-42.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. of NAACL-HLT 2019, Minneapolis, MN, USA, 2019, [Online]. Available: https://arxiv.org/abs/1810.04805.

[8] C. McAuley, J. Stewart, G. Siemens, and D. Cormier, The MOOC Model for Digital Practice, 2010, [Online]. Available: https://oerknowledgecloud.org/.

[9] I. Dagan, D. Glickman, and B. Magnini, “The PASCAL recognising textual entailment challenge,” in Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, U.K., 2005.

[10] P. Koehn, Statistical Machine Translation. Cambridge, U.K.: Cambridge University Press, 2010.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems 26 (NeurIPS 2013), 2013, pp. 3111-3119.