КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.912
IRSTI 16.31.61
https://doi.org/10.53364/24138614_2025_38_3_11

I. Ryabchenko[1], **R. Anayatova**[1], **G. Tulekova**[1], **A. Koshekov**[1*], **Y. Kuanov**[1]
[1]Civil Aviation Academy, Almaty, Republic of Kazakhstan

[*]*E-mail: a.k.koshekov@gmail.com*

MODELING SEMANTIC RELATIONSHIPS OF AVIATION TERMS: VECTOR SPACES AND LANGUAGE MODELS

Abstract. *The proposed article examines methods for modeling semantic relationships of aviation terms using the BERT and RoBERTa language models. The relevance of the study lies in the use of a pre-prepared and annotated corpus of aviation terms that align with international practice and are drawn from documents of international regulatory bodies. The developed language corpus provides the basis necessary for assessing the semantics of aviation terminology in the context of real aircraft operation. The research methodology involves fine-tuning language models trained on an aviation corpus of terms using cosine similarity, rank correlation, and cluster metrics of measurements. The experiments demonstrate the main differences between the two models in tracking synonyms, variability, and shifts in aviation discourse. The results of the study demonstrate that fine-tuning the models enhances their ability to cluster related terms, distinguish closely related but distinct concepts, and align the results with expert judgments. These results provide a methodological basis for the development of aviation terminology resources, enabling the application of lexicography transformer models and ontology construction.*

Keywords: *semantic proximity, aviation terminology, language models, corpus linguistics, transformers, embedding, natural language processing.*

Introduction.

Today, one of the primary tasks of natural language is modeling the semantic similarity of terms, and a substantial number of studies aim to address this problem. The problem of assessing the semantic similarity of terms arises in various practical applications, including automatic text translation, search for relevant information, intelligent search, ontology construction, and text analysis and processing for specific subject areas [1]. In a highly technical field and specialized industry like aviation, the accuracy of terminology is a crucial aspect, as the correctness of language structures directly determines professional international communication and document flow, which is closely related to ensuring flight safety.

Historically, the first approaches to modeling semantic relationships between terms were based on statistical methods, such as TF-IDF, which account for the frequency of a particular term without analyzing its content [2]. The next step in developing these methods is the use of distributed representations, such as Word2Vec and GloVe [3], which provide a representation of words in the form of vector spaces that capture semantic patterns through contextual co-occurrence in texts. However, such models assigned words a fixed vector, which did not allow for effective processing of polysemy and domain-specific meanings.

The next key step in the development of natural language processing is the use of contextualized language models, such as ELMo, BERT, and RoBERTa. These models are trained on a large language corpus using an attention function that provides context-sensitive embeddings to distinguish the synonymous meanings of terms [4]. These models yield fairly good results in natural language processing; however, their effective application in specialized areas depends on the adaptation of the corpus used, which is specific to each subject area. A feature of applications in the aviation industry is the complex technical terminology, multilingualism, and the use of various semantic changes depending on the content being processed [5]. The relevance of the topic is confirmed by the importance of accurate terminology modeling in ensuring the safe operation of aviation equipment, as well as lexicography and knowledge management. Existing research is primarily focused on the use of universal pre-trained models, which often fail to accurately reflect the semantic structure of aviation terminology [6, 7].

The aim of this study is to address the issue of utilizing language models in the aviation industry by developing and training a model on a new corpus of terminology directly related to aviation. The corpus of terms formed in this study comprises regulatory documents from international aviation authorities, technical manuals for aircraft operation, and training materials [8]. As part of the study, the authors manually annotated the corpus to establish semantic relationships, hierarchies between terms, and synonymous variations [9].

The main problem facing this study was the lack of a comprehensive corpus of terms with multilevel annotations. This complicates the comparison of language models, which negatively affects the development of lexicographic resources. This study aims to address this research gap by creating a specialized corpus of aviation terms from international regulatory documents, detailing the term annotation scheme, determining their quality, and comparing standard models with those trained on the developed language corpus.

As part of the study, the ability to reflect semantic proximity in aviation communication is assessed by fine-tuning the BERT and RoBERTa models on the collected corpus and comparing their performance with that of versions in the general subject area [10].

The objectives of this study are:

To create a corpus of aviation terminology and describe the methodology for its annotation;

To evaluate the effectiveness of the BERT and RoBETRa models for modeling the semantic similarity of aviation terms;

To compare the models using quantitative metrics (cosine similarity, rank correlation, clustering indices);

To visualize the semantic space using dimensionality reduction methods for accurate interpretation of terms.

The scientific novelty of the study lies in the combination of transformer approaches with the use of a subject-oriented annotated language corpus, which enables the utilization of the semantic behavior of terminology in a real professional context. The practical significance lies in the application of the results in the construction of an aviation thesaurus, ontologies, and the development of an intelligent linguistic support system for aviation industry specialists.

Materials and methods.

Formation of the aviation terminology corpus

To conduct the study and enhance the reliability of the experimental results, the authors created a specialized corpus of aviation terms. The material for the corpus was collected from authoritative industry sources, including ICAO Annexes, EASA and FAA regulations and guidelines, IATA operational standards, Airbus and Boeing aircraft operating manuals, aviation textbooks, and scientific articles. The corpus of words comprises 100,000 unique tokens, ensuring comprehensive coverage of all key subdomains within the aviation industry. The tokens include terms from the time period 2000 to 2025. The breakdown of tokens by subdomains is presented in Table 1.

Table 1 – Corpus statistics by aviation subdomains

Subdomain	Documents	Tokens	% of corpus	Unique terms

Navigations	420	20 000	20	2 600
Flight operations	350	18 000	18	2 200
Safety and emergency	310	16 000	16	2 000
Maintenance and emergency	300	15 000	15	1 900
Airworthiness and regulations	240	12 000	12	1 500
Aircraft structures and parts	210	10 000	10	1 300
Air traffic control and communication	170	9 000	9	1 100
Total	2 000	100 000	100	12 600

The terminology set was selected to be well-balanced across all aviation subdomains. This proportional division reflects the practical importance of each area in aviation communication. The corpus also includes additional terminology, such as maintenance, regulatory language, and phraseology used in air traffic control, which are also important for ensuring flight safety.

The pre-processing process includes tokenization, text cleaning, and normalization, after which the text is annotated at multiple levels. The first level includes morphological markup to ensure the participatory nature of the information and lemmatization. The second level includes terminological markup with the assignment of labels according to aviation subdomains (navigation, safety, etc.). The third level introduces semantic relations such as synonymy, hypernymy, and the part-whole relation, which allows the corpus to reflect the full complexity of professional aviation terminology. The dataset is verified by aviation linguists to ensure terminological accuracy. The process flow diagram for creating and annotating the aviation terminology language corpus is shown in Figure 1.

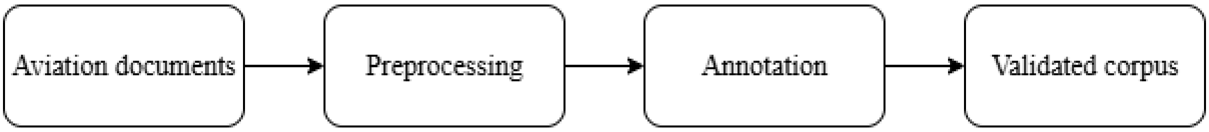


Figure 1 – Scheme of corpus construction and annotation

To demonstrate the term annotation protocol and ensure transparency, Table 2 presents examples of term annotations. Each entry represents the term's canonical form, subdomain, semantic relationship type, representation variants, and example of the term's use in a sentence. This display format demonstrates the multilevel annotation scheme and the encoding of linguistic and semantic information for model training and evaluation.

Table 2 – Examples of annotated aviation terms

Term	Canonical form	Subdomain	Relation	Variants	Example
Briefing	Briefing	Safety	Synonymy	Preflight briefing	The crew conducted a safety briefing prior to engine start
Altitude	Altitude	Navigation	Hypernymy	Flight level	Maintain altitude 3500 feet until

					cleared for approach
Autopilot	Autopilot	Operations	Associative	Flight direction	The autopilot was engaged during climb to reduce pilot workload
Fuselage	Fuselage	Structures	Meronymy	Airframe	The fuselage inspection revealed minor structural damage
GPWS	Ground proximity warning system	Safety	Abbreviation	-	The GPWS alert required an immediate go-around

Table 2 shows that the annotation incorporates both the lexical and semantic aspects of aviation terminology. In the dataset, "briefing" is associated with synonyms in the safety subdomain, while "fuselage," with its part-whole relationship, belongs to the structure subdomain. The abbreviations and multiword variants of terms are explicitly recorded to ensure the model's ability to clearly recognize equivalences. Examples of term usage in sentences provide a contextual basis, and are necessary for expert verification.

Model fine-tubing

Two-stage model fine-tuning and training parameters for universal models were adapted for aviation crops. The first optimization step is domain adaptation using masked language modeling (MLM), which helps models learn the distributional semantics of tokens to predict a randomly masked term from the context:

$$\mathcal{L}_{\mathrm{MLM}} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathrm{P}(\mathrm{w_i}|\widehat{\mathrm{w}}_{\mathrm{i}}, \theta), \tag{1}$$

where $\mathrm{w_i}$ – original token, $\widehat{\mathrm{w}}_{\mathrm{i}}$ – masked version of the token, θ – set of model parameters.

The second step is contrastive learning with triplet loss, which generates a triplet of "anchor-positive-negative" terms based on the annotated complex. This configuration ensures a semantic relationship between the anchor and positive terms while assigning negative terms to a different subdomain. This procedure helps to separate related but conceptually different terms, increasing the robustness of term embedding in professional contexts.

For the MLM, dynamic masking with a probability of 15% was used to ensure a variety of training scenarios. The retraining process was controlled using validation-perplexity-based stopping. Model evaluation was conducted using the MLM and MLM+triplet configurations, allowing for the quantitative evaluation of each stage to be examined. The training settings for the experiment included the following parameters: learning rate MLM ($5x10^{-5}$), learning rate triplet ($2x10^{-5}$), batch size – 32, epochs – 20, warmup ratio – 10%, weight decay – 0.01, optimizer – AdamW, validation – 10%, random seed – 42 and hardware – 24 gb vram.

These parameters ensure a balance between the computational efficiency and convergence. The stability of fine-tuning was ensured using warmup learning and parameter weight reduction, while reproducibility was achieved using a fixed initial value. Triplet loss improves the separation of semantically related terms, as reflected in the ablation results.

Semantic similarity estimation

To effectively evaluate the relationships between aviation terms, embeddings were extracted from the proposed models using the SentenceTransformers framework. The embeddings from the models have dimensions of 384 for BERT and 768 for RoBERTa, which provides a detailed representation of the semantic space. The main similarity measure used in this paper is cosine similarity, which helps estimate the proximity of term vectors in space without relying on absolute values. The similarity between two terms t_1 and t_2 is determined using the formula:

$$sim(t_1, t_2) = \frac{v_{t1} \cdot v_{t2}}{\|v_{t1}\| \cdot \|v_{t2}\|}, \quad (2)$$

where the numerator is presented as a scalar product of vectors and the denominator normalizes the values by the Euclidean norm.

The convergence coefficient ranges from -1 to 1, where a higher value indicates stronger semantic similarity. To illustrate this approach, Figure 2 was prepared, which shows an example of calculating similarity using cosine similarity for a pair of aviation terms. The term "vectors" is presented in a two-dimensional coordinate space, where the angle between the vectors corresponds to the strength of semantic similarity.

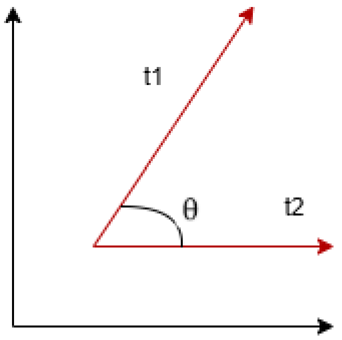
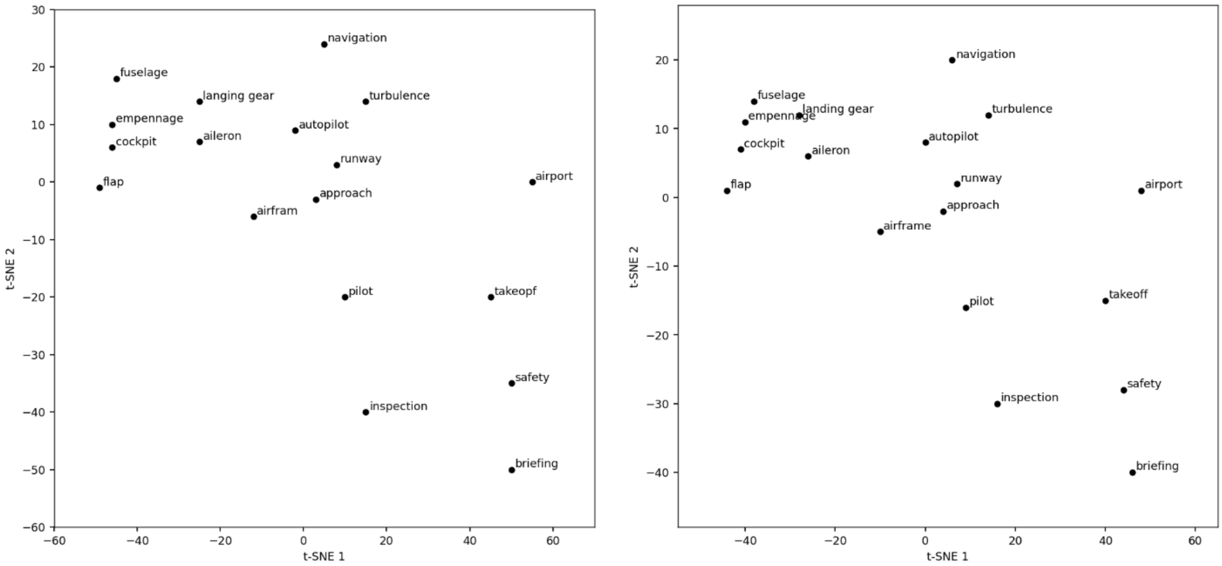


Figure 2 – Example of cosine similarity calculation between aviation term pairs

Semantic evaluations were conducted using internal and external tests to ensure the quantitative reliability and professional validity. For internal evaluation, a dataset of 500 paired terms was created and evenly balanced across four relationship types: synonymy, hypernymy, meromyny (part-whole), and disjoint (contrast) pairs. Each pair of terms was rated by aviation linguists and industry experts and assigned a score from 0 to 4. The annotators' agreement was estimated at 0.82, indicating a high level of agreement. Each model was evaluated using cosine similarity between term vectors, and the correlation between expert assessments was measured using Spearman's rho. Robustness was ensured by resampling 1000 times using the bootstrap method (95% confidence interval). Differences between the models were assessed using Fisher's z-transform. The clustering quality was assessed based on the annotated subdomains. Silhoutte and Davies-Bouldin indices were calculated for this assessment. Stochastic effects were accounted for by averaging 10 random initial values. External tests addressed two practical issues. The first was the automatic matching of terms with external thesuri using precision and recall assessments. The second was predicting the correct assignment of meanings for polysemantic terms assessed using expert annotations. In addition to the numerical model indicators, experts manually verified the model's results. This multilevel approach ensured the stochastic and linguistic validity of the results.

Visualization of semantic spaces

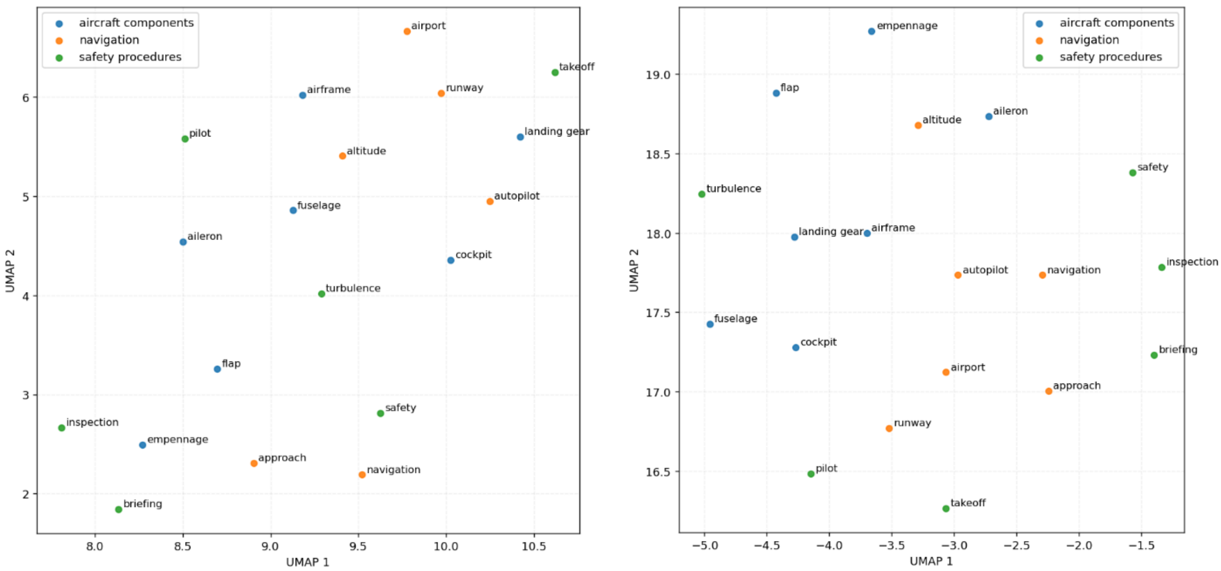
Domain-adapted embeddings structure aviation terminology by projecting multidimensional vectors onto two dimensions using t-SNE and UMAP. T-SNE is used to identify compact groups of terms that are closely related. The stochastic nature of the model is controlled by introducing a fixed seed, the perplexity is tuned to a moderate range to balance local granularity and global readability, and the number of iterations is set high enough to ensure convergence. The resulting t-SNE maps from the study are presented in Figure 3, where semantically consistent clusters (e.g., aircraft components, navigation, safety) are highlighted, and subtle differences within the clusters themselves (e.g., control surfaces and structural nodes) are often obscured in quantitative values.



a) BERT b) RoBERTa

Figure 3 – t-SNE projection of aviation embeddings

In addition, UMAP is used for analysis to examine the local and global structure of diversity, employing a graph approach that ensures the preservation of inter-cluster geometry. Cosine distance is used for the angular nature of similarity in space, which helps to avoid database overflow and preserve intra-class relationships. The projections presented in Figure 4 show a consistent macrostructure for both models, making the inter-cluster separation more obvious compared to t-SNE, which helps capture category leakage (i.e., the shift of exploitable terms towards safety).



a) BERT b) RoBERTa

Figure 4 – UMAP projection of aviation embeddings

These plots provide complementary evidence across domains. When fine-tuned on a corpus of aviation terms, BERT produces tight, discrete clusters with sharp boundaries, which is useful in tasks that require clear class separation. RoBERTa has smoother transitions between adjacent subdomains, which is useful in cases where graded similarity is needed. Both models use a fixed seed, hyperparameters, and term set, which is necessary to allow direct comparisons between them. The limitation of 2D projections is that they discard information and introduce layout artifacts, which is addressed in this paper by using quantitative metrics that provide openness and interpretability in the geometry of aviation terminology.

Evaluation protocol

The semantic similarity of terms is assessed using a combination of quantitative metrics and qualitative expert assessment by linguistic teachers. The verification process is designed to assess the mathematical consistency of the results and their compliance with professional linguistic terminology in the aviation industry. The quantitative assessment of the vector proximity measure was determined using cosine similarity. A reference dataset is used to verify the calculated values. This dataset comprises carefully selected pairs of terms that have been independently annotated by linguists and experts from aviation enterprises. The correlation between similarity estimates obtained using the models and expert assessments is calculated using Spearman's rank correlation coefficient to assess the monotonicity of the relationship between similarities.

In addition to pairwise similarity, the quality of clustering is investigated to determine the quality of reflection of the high-level structure of aviation terminology. The silhouette index and Davis-Bouldin are used to evaluate grouping into semantic clusters, which provides a representation of intra-cluster coherence and inter-cluster separation. These metrics complement the overall assessment, which is based on similarity, tracking, subgroup identification consistency, and domain specificity.

Linguistic experts assessed the representativeness of a subset of the model results. This expert assessment is crucial for confirming the practical significance of the results, as some semantic relationships may not be fully reflected when evaluating statistical indicators. The assessment of the model's ability to accurately represent aviation terminology is achieved by integrating clustering indicators, expert assessment, and a protocol to ensure the multidimensionality and reproducibility of the assessment. Such an integrated approach ensures methodological sustainability and practical significance of the results of lexicographic and terminological applications in the aviation industry.

Results and discussion.

Pairwise semantic similarity

The key method for assessing the effectiveness of semantic models is the comparison of pairwise similarity scores between aviation terms. The cosine convergence coefficient is used to assess the degree of correspondence of the selected term pairs in the semantic space. The term pairs are selected taking into account typical relationships adopted in the aviation industry: near-synonyms (aircraft-plane), functional relatedness (autopilot-flight director), contextual similarity (safety-inspection), and semantic similarities (cockpit-fuselage). Table 3 presents the similarity scores for the term pairs described above. The results are presented for the universal versions of BERT and RoBERTa, as well as optimized versions of these models trained on the aviation terminology corpus.

Table 3 – Cosine similarity scores for selected aviation term pairs under general and fine-tuned models

Term pair	General BERT	Fine-tuned BERT	General RoBERTa	Fine-tuned RoBERTa
Aircraft-plane	0.62	0.89	0.65	0.91
Runway-airstrip	0.58	0.87	0.60	0.85
Autopilot-flight director	0.49	0.76	0.55	0.82
Turbulence-disturbance	0.44	0.72	0.47	0.75
Cokpit-fuselage	0.33	0.64	0.35	0.61
Safety-inspection	0.39	0.70	0.42	0.73

The obtained results demonstrate that fine-tuning improves the consistency of vectors and expert judgment in the aviation industry. For example, for the pair of terms “airstrip” and “runway”, the similarity score after fine-tuning the BERT model increased from 0.58 to 0.87, and for the RoBERTa model from 0.60 to 0.85, which means that they are used synonymously. Additionally, terms with functional or procedural overlap also increased the convergence rate after fine-tuning the models, a finding confirmed by expert judgment.

An interesting difference is demonstrated between the two models on pairs containing functional differences. For example, the RoBERTa model shows high stability for similar pairs, indicating high sensitivity to graded similarity. In contrast, the BERT model shows a sharp increase in convergence on synonymous words, indicating a strength in forming discrete equivalence classes. Overall, when analyzing pairwise convergence, we can conclude that both models improve their performance after fine-tuning, with complementary strengths.

Table 4 presents the performance differences between the methods (TF-IDF, Word2Vec, GloVe, and FastText) and the transformer-based models (BERT and RoBERTa) in the baseline settings and with fine-tuning for aviation terminology. The results were presented as Spearman correlations and clustering metrics.

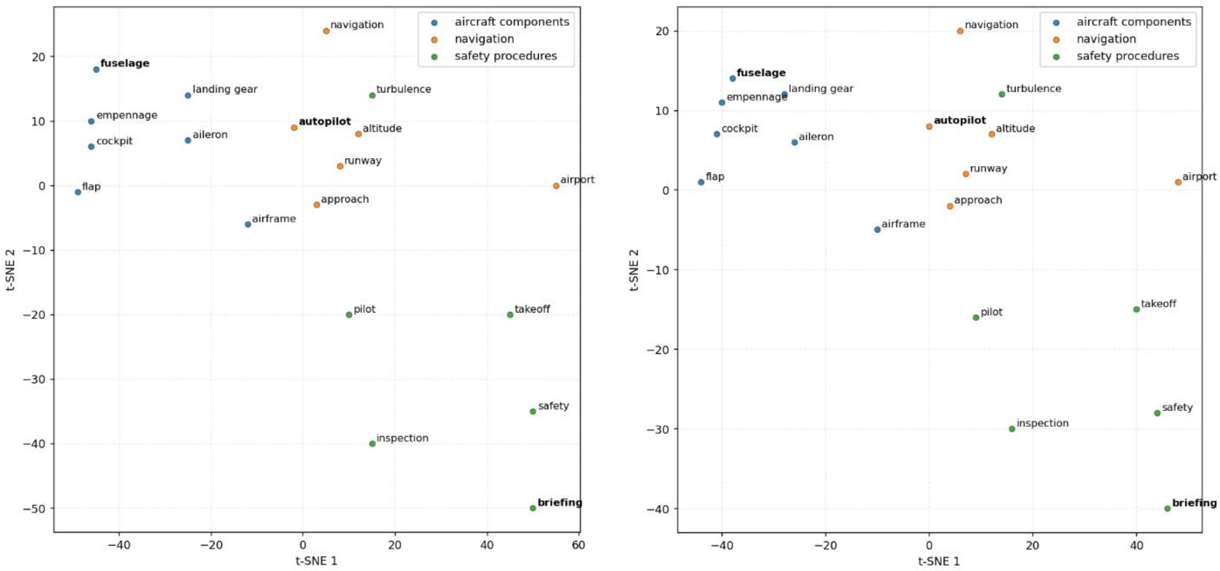
Table 4 – Comparison of baseline and transformer models

Model	Spearman	Silhoutte	Davies-Bouldin
TF-IDF	0.41	0.28	1.92
Word2Vec	0.53	0.36	1.65
GloVe	0.55	0.38	1.59
FastText	0.59	0.42	1.48
BERT	0.62	0.44	1.37
BERT (aviation)	0.83	0.61	1.01
RoBERTa	0.65	0.46	1.35
RoBERTa (aviation)	0.81	0.58	1.08

The results demonstrated three key findings. First, both transformer models outperformed the classical baseline models, confirming the superiority of the contextual models. Fine-tuning enables the capture of subtle semantic relationships. Second, the models are domain-adaptable; fine-tuned models demonstrate a 0.2-point increase in Spearman coefficient compared to their generic counterparts, and two other metrics demonstrate effectiveness, demonstrating the formation of coherent clusters. Third, both transformer architectures have mutually exclusive advantages: fine-tuned BERT produces clear and discrete clusters, enabling effective ontologies and dictionary development, whereas RoBERTa produces smooth semantic gradients between terms, which facilitates term similarity search.

Visualization of semantic spaces

To provide a qualitative check of the structuring of aviation terminology, the authors examined Figure 3 by dividing it into subdomains using a color scheme to better demonstrate the differences. Adaptation to the domains yields clusters that align with the experts’ expectations. In Figure 5(a), the clusters have a denser and more discrete grouping structure. This geometry is consistent with the BERT model’s tendency to have clear decision boundaries, often related to tasks such as terminological labeling and dictionary meaning assignment.



a) BERT model b) RoBERTa model

Figure 5 – Scatter plot for both models with consistent color subdomains coding

In Figure 5(b), the clusters remain stable and show smooth gradients along their boundaries, for example the term "autopilot" in this model is closer to the operational sample, while maintaining its distance from the structural lexicon subdomain, while "briefing" still belongs to the safety cluster, but shows a move towards operational, due to its use in pre-flight procedures. Such smooth transitions make the RoBERTa model a valuable tool for synonym search and expansion of the aviation lexicon, where nuances of proximity are more effective than rigid partitioning.

Using a uniform color scheme and term lists allows for a visual comparison of true differences in term embeddings. Random seeds and identical dimensionality reduction settings are fixed for both models to ensure reproducibility. Two-dimensional projections compress information and may distort long-term relationships, but the observed patterns confirm the results obtained in the previous two subsections. Based on the analysis, we can conclude that fine-tuning enhances intra-cluster connectivity and improves inter-cluster separation. The BERT model ensures separability, while RoBERTa captures semantic gradients.

Expert validation

The previously proposed quantitative metrics provide mechanisms for assessing semantic models, but for specific domains such as aviation, it is necessary to rely on expert understanding of terminology. To achieve this, a group of linguists and representatives of aviation enterprises was involved in the study, who examined subsets of the model results and compared them with semantic expectations. The experts confirmed that the fine-tuned model better matched professional interpretations of aviation terms. The generic model often associates terms with every day or general language, which often differs from the specialized use of these terms in aviation. In contrast, the fine-tuned proposed models effectively reinforced context-dependent and aviation-specific relationships. A brief example of the comparison between expert assessments and the base model's judgments is presented in Table 5.

Table 4 – Case-study examples of expert validation

Term	Closet term in general model	Closet term in fine-tuned model	Expert judgement	Expert comment (general model/fine-tuned model
Briefing	communication	inspection	Safety context	General talk/Safety

Altitude	height	approach	Navigation	Everyday synonym/Aviation specific role
Autopilot	robotics	Flight director	Operations	Robotics/Cockpit system
Airframe	Construction	Fuselage	Aircraft components	Construction/Structural vocabulary
Takeoff	Departure	Runway	Flight operation	Travel synonym/Operational aviation context
Turbulence	chaos	safety	Safety procedures	Abstract notion/Procedural and safety term

These examples demonstrate the importance of creating domain-specific term corpora. Basic models often linked technical aviation terms to a broad everyday context. Fine-tuned models align terms with professional usage, ensuring that clusters and convergence rates match authentic linguistic practice. Conformity to expert assessment demonstrates the methodological soundness and practical value of this approach, especially in cases of lexicographic databases, training materials, and safety documentation.

Implications and limitations

The results obtained in this study demonstrate the strengths and weaknesses of applying the transformer-based models to the aviation terminology domain. Fine-tuning the models shows improvements in semantic similarity, clustering quality, and expert judgment. These achievements demonstrate the practical importance of adapting models to a specific domain, which helps avoid distorting the meaning of technical terms when comparing them with everyday synonyms or unrelated concepts. The creation of a specialized aviation corpus helps to create a reliable terminology resource for curriculum improvement and intelligent search in the aviation context.

In addition to the advantages, this study has several limitations in its results. First, the corpus used for fine-tuning the models remains small in size, which limits the representativeness of the training material and increases the risk of overfitting the model on specific subsets of aviation subdomains. Second, the annotation process is limited by the resources of the academic institution and does not reach the level of coverage typical of large-scale corpus linguistics projects. Another methodological limitation is the use of two-dimensional projection as an interpretation tool, which compress multivariate relationships and may obscure the boundaries of some clusters. Therefore, the insights gained from scatterplots should always be considered in conjunction with quantitative metrics. Although fine-tuning resulted in measurable improvements, the results exhibit model-specific biases: BERT favors crisp and discrete clusters, leading to the simplification of semantic continua, while the RoBERTa model exhibits smooth gradients that may blur distinctions critical for aviation safety terminology. These limitations underscore the importance of careful model selection and the ability to combine models based on the task they are intended to perform.

Based on the limitations described, it can be concluded that this study, while developing a specific subject area and offering a replicable assessment system, represents only the first step in creating a comprehensive dictionary of aviation terms. Future work involves expanding the corpus of terms to include multilingual data reflecting the identity of each state, as well as participants in international organizations. Additionally, it involves introducing advanced assessment methods that take into account pragmatic and discursive phenomena beyond pairwise similarity or clustering.

Conclusion.

The main objective of this study was to evaluate the transformer-based language models for aviation terminology by creating a specialized corpus and fine-tuning the BERT and RoBERTa models. The obtained results demonstrate that domain-specific adaptation increases the semantic accuracy of vector representations, improves the quality of the clustering process, and ensures accurate compliance with expert assessments. These results demonstrate the need for a specialized corpus in the aviation industry and indicate the potential application in databases, training materials, and decision support systems. At the same time, the study reveals critical limitations: to date, the corpus does not have a high level of scale and depth of annotation, the methods of visualization of results are useful for analytics, but certain structural properties of the embedding space cannot be taken from it, and specific distortions characteristic of these models indicate the need for a hybrid solution for processing aviation linguistics.

This study represents the first step in creating a comprehensive dictionary of aviation terms. The primary contribution of this study is to confirm the hypothesis that we need to build our own language corpus and fine-tune it to the specific needs of the aviation industry. For aviation linguistics and applied natural language processing, the results underscore the need for further research on developing a large, multilingual, and systematically annotated language corpus, as well as hybrid methods for evaluating model performance. With continued efforts, it is possible to provide reliable, reproducible, and practically applicable tools for terminology, lexicography, and security of communication.

Funding.

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23490578) «Development of linguistic support and mobile application for a Kazakh-English-Russian thesaurus-type dictionary of modern aviation terms regulated by ICAO, EASA, IATA standards».

References

1. Jiang, J., & Wang, H. H. (2020). Application intelligent search and recommendation system based on speech recognition technology. International Journal of Speech Technology, 24(1), 23–30. https://doi.org/10.1007/s10772-020-09703-0.

2. Zhuohao, W., Dong, W., & Qing, L. (2021). Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF. Chinese Journal of Electronics, 30(4), 652–657. https://doi.org/10.1049/cje.2021.05.007

3. Al-Aidaroos, A. S., & Bamzahem, S. M. (2023). The Impact of GloVe and Word2Vec Word-Embedding Technologies on Bug Localization with Convolutional Neural Network. International Journal of Science and Engineering Applications, 108–111. https://doi.org/10.7753/ijsea1201.1035

4. Wada, S., Takeda, T., Okada, K., Manabe, S., Konishi, S., Kamohara, J., & Matsumura, Y. (2024). Oversampling effect in pretraining for bidirectional encoder representations from transformers (BERT) to localize medical BERT and enhance biomedical BERT. Artificial Intelligence In Medicine, 153, 102889. https://doi.org/10.1016/j.artmed.2024.102889

5. Tang, H., Kamei, S., & Morimoto, Y. (2023). Data Augmentation Methods for Enhancing Robustness in Text Classification Tasks. Algorithms, 16(1), 59. https://doi.org/10.3390/a16010059

6. Gandour, J. (2002). Interdisciplinary Approaches to Language Processing (review). Language, 78(2), 359. https://doi.org/10.1353/lan.2002.0081

7. Muraki, E. J., Feddema, C. Y., Pexman, P. M., & Reggin, L. D. (2023). The Development of Abstract Word Meanings. Journal of Child Language, 52(1), 195–207. https://doi.org/10.1017/s0305000923000569

8. Zhao, L., Fang, J., & Gao, W. (2024). Optimizing Large Language Models on Multi-Core CPUs: A Case Study of the BERT Model. Applied Sciences, 14(6), 2364. https://doi.org/10.3390/app14062364

9. Moe, R. (2010). Compiling dictionaries using semantic domains. Lexikos, 13(1). https://doi.org/10.4314/lex.v13i1.51389

10. Shu, X. (2024). BERT and RoBERTa for Sarcasm Detection: Optimizing Performance through Advanced Fine-tuning. Applied and Computational Engineering, 97(1), 1–11. https://doi.org/10.54254/2755-2721/97/20241354

ТЕРМИНДЕРДІҢ СЕМАНТИКАЛЫҚ БАЙЛАНЫСТАРЫН МОДЕЛЬДЕУ: ВЕКТОРЛЫҚ КЕҢІСТІКТЕР ЖӘНЕ ТІЛДІК МОДЕЛЬДЕР

***Аңдатпа**. Ұсынылған мақалада авиациялық терминдердің семантикалық қатынастарын модельдеудің әдістері BERT және RoBERTa тілдік модельдерін қолдану арқылы қарастырылады. Зерттеудің өзектілігі алдын ала дайындалған және аннотацияланған авиациялық терминдер корпустарын пайдаланумен айқындалады, олар халықаралық тәжірибеге сәйкес келеді және халықаралық реттеуші органдардың құжаттарынан алынған. Құрастырылған тілдік корпус ұшақтарды нақты пайдалану жағдайында авиациялық терминология семантикасын бағалау үшін қажетті негізді қамтамасыз етеді. Зерттеу әдіснамасы авиациялық терминдер корпусында алдын ала үйретілген тілдік модельдерді қайта баптауды (fine-tuning) қамтиды және ол косинустық ұқсастық, рангілік корреляция мен кластерлік метрикаларды өлшеуге негізделген. Эксперимент нәтижелері екі модельдің синонимдерді қадағалау, вариативтілік пен авиациялық дискурстағы мағыналық ығысудағы негізгі айырмашылықтарын көрсетті. Зерттеу нәтижелері модельдерді қайта баптау олардың байланысты терминдерді топтастыру қабілетін арттыратынын, өзара ұқсас, бірақ әртүрлі ұғымдарды ажырата алатынын және нәтижелерді сарапшылардың бағаларымен сәйкестендіретінін дәлелдейді. Бұл нәтижелер авиациялық терминологиялық ресурстарды дамытуға әдістемелік негіз ұсынады, трансформер модельдерін лексикографияда және онтология құруда қолдануға мүмкіндік береді.*

***Түйін сөздер**: семантикалық жақындық, авиациялық терминология, тілдік модельдер, корпус лингвистикасы, трансформерлер, эмбеддинг, табиғи тілдерді өңдеу.*

МОДЕЛИРОВАНИЕ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ТЕРМИНОВ: ВЕКТОРНЫЕ ПРОСТРАНСТВА И МОДЕЛИ ЯЗЫКА

***Аннотация.** Предлагаемая статья рассматривает методы моделирования семантических отношений авиационных терминов с использованием языковых моделей BERT и RoBERTa. Актуальность исследования заключается в применении заранее подготовленного и аннотированного корпуса авиационных терминов, который соответствует международной практике и сформирован на основе документов международных регулирующих организаций. Разработанный языковой корпус обеспечивает необходимую основу для оценки семантики авиационной терминологии в контексте реальной эксплуатации воздушных судов. Методология исследования включает дообучение (fine-tuning) языковых моделей на корпусе авиационных терминов с использованием косинусного сходства, ранговой корреляции и кластерных метрик. Эксперименты продемонстрировали основные различия между двумя моделями в отслеживании синонимов, вариативности и семантических сдвигов в авиационном дискурсе. Результаты исследования показали, что дообучение моделей повышает их способность кластеризовать связанные термины, различать близкие, но отличающиеся понятия, а также согласовывать результаты с экспертными оценками. Полученные данные обеспечивают методологическую основу для разработки ресурсов авиационной терминологии, что позволяет применять трансформерные модели в лексикографии и построении онтологий.*

Ключевые слова: *семантическая близость, авиационная терминология, языковые модели, корпусная лингвистика, трансформеры, эмбеддинг, обработка естественного языка.*

Information about the authors

Ryabchenko Inna	Research Fellow, Civil Aviation Academy, Almaty, Kazakhstan E-mail: i.ryabchenko@agakaz.kz
Anayatova Raziyam	PhD, Head of the Department of Aviation English, Civil Aviation Academy, Almaty, Kazakhstan, E-mail: r.anayatova@agakaz.kz
Tulekova Gulnaz	PhD, Professor, Department of Aviation English, Civil Aviation Academy, Almaty, Kazakhstan, g.tulekova@agakaz.kz
Koshekov Abay	PhD, Associate Professor, Department of Aviation Engineering and Technology, Civil Aviation Academy, Almaty, Kazakhstan, E-mail: a.k.koshekov@gmail.com
Kuanov Yerkanat	Research Fellow, Civil Aviation Academy, Almaty, Kazakhstan E-mail: e.kuanov@agakaz.kz

Авторлар туралы мәлімет

Рябченко Инна	Азаматтық авиация академиясының ғылыми қызметкері, Алматы, Қазақстан, E-mail: i.ryabchenko@agakaz.kz
Анаятова Разиям	PhD, Азаматтық авиация академиясының авиациялық ағылшын тілі кафедрасының меңгерушісі, Алматы, Қазақстан, E-mail: r.anayatova@agakaz.kz
Тулекова Гульназ	PhD, профессор, авиациялық ағылшын тілі кафедрасы, Азаматтық авиация академиясы, Алматы, Қазақстан, g.tulekova@agakaz.kz
Кошеков Абай	PhD, азаматтық авиация академиясының авиациялық техника және технологиялар кафедрасының доценті, Алматы, ҚазақстанE-mail: a.k.koshekov@gmail.com
Куанов Ерканат	Азаматтық авиация академиясының ғылыми қызметкері, Алматы, Қазақстан, E-mail: e.kuanov@agakaz.kz

Сведение об авторах

Рябченко Инна	Научный сотрудник Академии гражданской авиации, Алматы, Казахстан, E-mail: i.ryabchenko@agakaz.kz
Анаятова Разиям	PhD, заведующий кафедры «Авиационный английский язык» Академии гражданской авиации, Алматы, Казахстан, E-mail: r.anayatova@agakaz.kz
Тулекова Гульназ	К.ф.н., профессор кафедры «Авиационный английский язык» Академии гражданской авиации, Алматы, Казахстан, g.tulekova@agakaz.kz
Кошеков Абай	PhD, ассоциированный профессор кафедры «Авиационная техника и технологии» Академии гражданской авиации, Алматы, Казахстан E-mail: a.k.koshekov@gmail.com
Куанов Ерканат	Научный сотрудник Академии гражданской авиации, Алматы, Казахстан E-mail: e.kuanov@agakaz.kz