UDC 004.855.6
IRSTI 20.19.19

https://doi.org/10.55452/1998-6688-2025-22-4-227-243

[1]*Mukhsimbayev B.,
PhD student, ORCID ID: 0009-0008-4606-3628,
*e-mail: b.mukhsimbaev@kbtu.kz
[1]Pak A.,
PhD, Professor, ORCID ID: 0000-0002-8685-9355,
e-mail: a.pak@kbtu.kz
[1]Kuralbayev A.,
PhD student, ORCID ID: 0009-0001-0811-5385,
e-mail: a.kuralbaev@kbtu.kz

[1]Kazakh-British Technical University, Almaty, Kazakhstan

A COMPUTATIONAL PIPELINE FOR LEXICAL AND THEMATIC ANALYSIS OF THE CODE OF ADMINISTRATIVE OFFENSES OF THE REPUBLIC OF KAZAKHSTAN

Abstract

This study introduces a computational pipeline for the automated linguistic and structural analysis of legal texts, applied to the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK, K1400000235). The proposed workflow integrates data collection, text preprocessing, tokenization, keyword extraction, semantic clustering, and visualization using natural language processing (NLP) and statistical techniques implemented in Python. The pipeline unites lexical, thematic, and quantitative linguistic analyses into a coherent sequence that enables the identification of frequency distributions, semantic fields, and latent topics across the hierarchical structure of the Code (sections, chapters, and articles). The analysis of the CAO RK corpus revealed several distinctive linguistic patterns: a dominance of sanction and responsibility-related vocabulary (штраф, ответственность, правонарушение), high lexical density in chapters regulating economic and procedural offenses, and concentrated thematic clusters reflecting the normative-punitive orientation of administrative law. Visualization techniques such as frequency histograms, thematic heatmaps, and topic maps illustrate the potential of the pipeline for exploring legislative language quantitatively. Overall, the framework establishes a scalable foundation for comparative legal linguistics, automated legislative monitoring, and the modernization of legal analytics in Kazakhstan.

Keywords: administrative law, legal text analysis, natural language processing, computational legal linguistics, frequency analysis, topic modeling, legal informatics.

Introduction

Legal texts are among the most structured and linguistically formal types of documents, embodying both the conceptual logic and the institutional framework of law. As corpus linguistics continues to evolve globally, the creation of domain-specific linguistic corpora – including legal and administrative texts – has become a crucial research direction [1]. Such corpora not only preserve linguistic authenticity but also enable the systematic study of how institutional discourse reflects socio-legal priorities.

The Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK) is a core legislative act regulating administrative liability, sanctions, and procedural mechanisms within the national legal system. Comparative analyses of administrative liability across Commonwealth of Independent States (CIS) countries underscore Kazakhstan's distinctive legal structure and its emphasis on detailed qualifying characteristics [13].

However, quantitative or computational linguistic approaches to Kazakhstan's legal texts remain largely unexplored. The absence of large-scale empirical analyses of lexical and thematic regularities in the CAO RK creates a methodological gap between traditional legal scholarship and modern computational linguistics.

To address this gap, the present study applies computational linguistics and natural language processing (NLP) techniques to the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), offering an empirical perspective on its lexical and thematic organization. Specifically, the study introduces a unified computational pipeline that integrates lexical, thematic, and quantitative linguistic analyses within a single workflow.

Automated analysis of legislative texts and amendments in the Republic of Kazakhstan offers new possibilities for identifying thematic and lexical trends that reflect shifts in state policy priorities. Such an approach reduces the subjectivity of legal interpretation, enhances the transparency of normative evolution, and provides a foundation for more evidence-based interaction among legislators, researchers, and civil society.

By integrating computational methods with legal doctrine, it becomes possible to move beyond descriptive interpretation toward data-driven legal analytics that objectively trace how legal language changes over time and across domains. Recent works employing large language models for empirical legal studies demonstrate the growing potential of AI-assisted thematic and linguistic analysis within judicial and legislative corpora [14].

A preliminary quantitative analysis of the Code of Administrative Offenses of the Republic of Kazakhstan reveals a high lexical concentration of sanction-related and responsibility-oriented vocabulary, particularly within sections regulating economic and procedural offenses. This pattern underscores the normative-punitive orientation of administrative law and its focus on enforcement mechanisms rather than judicial adjudication. Thematic clustering further highlights distinct semantic zones within the Code, where linguistic density corresponds to institutional priorities such as governance, entrepreneurship regulation, and procedural enforcement.

These findings demonstrate the potential of computational linguistics to uncover structural and conceptual patterns that are not easily visible through traditional legal interpretation.

Specifically, the study introduces a unified computational pipeline that integrates lexical, thematic, and quantitative linguistic analyses within a single workflow. Applied to the CAO RK corpus, the pipeline integrates methods from NLP and legal informatics to:

1. Quantify lexical frequency and distribution of key legal terms;
2. Detect semantic clusters and thematic patterns through topic modeling;
3. Examine the variation of linguistic features across the hierarchical structure of the Code (sections, chapters, and articles).

The main contributions of this work are as follows:

1. Development of an end-to-end computational pipeline for the collection, preprocessing, and analysis of legislative texts in a reproducible manner;
2. Identification of lexical and thematic regularities in the CAO RK, including the predominance of sanction-related and responsibility-oriented vocabulary;
3. Discovery of structural and topical hierarchies within the Code through quantitative modeling and visualization (e.g., heatmaps and topic clustering);
4. Demonstration of how computational linguistic analysis can support legislative transparency, policy monitoring, and comparative studies of Kazakhstan's legal language.

Together, these contributions provide both a methodological innovation and new empirical insights into the structure and evolution of Kazakhstan's administrative law.

Related Works

Numerous linguistic and legal studies have examined the textual and terminological dimensions of Kazakhstan's legislation – including analyses of translation accuracy, linguistic expertise, and terminology standardization [2, 5, 9]. However, these works primarily focus on bilingual authenticity

and linguistic interpretation, while corpus-based and statistical approaches to the structure of Kazakhstan's codes remain limited.

Frameworks such as the Leipzig Corpus Miner illustrate how qualitative legal analysis can be enhanced through text-mining architectures that combine close and distant reading – an approach conceptually aligned with our computational pipeline for analyzing legislative corpora [16]. Recent advances in legal natural language processing have demonstrated how machine learning and large language models can assist in thematic and structural segmentation of legal texts, suggesting promising directions for applying similar techniques to normative corpora [14, 15].

Recent developments in computational linguistics for the Kazakh language – such as the creation of the KazQAD open-domain QA dataset, the KazSAnDRA sentiment analysis dataset, the Aligned Kazakh–Russian Criminal Corpus, large-scale instruction-tuning datasets for low-resource governance and legal domains, and corpus-driven approaches to crime-related event extraction for low-resource languages – have significantly expanded the available linguistic resources for Kazakh [6–8, 11, 12]. Nevertheless, most of these initiatives target general-domain or multilingual NLP rather than formal legislative corpora.

In parallel, legal scholars emphasize the ongoing challenges of ensuring authenticity and linguistic precision in Kazakh–Russian legal texts [2, 4, 9, 10]. These works underline the need for standardized terminology and systematic linguistic evaluation – issues that computational methods can help address by providing quantitative diagnostics of translation consistency and lexical coherence across codes.

Automated analysis of legislative texts and amendments in the Republic of Kazakhstan offers new possibilities for identifying thematic and lexical trends that reflect shifts in state policy priorities. Such an approach reduces the subjectivity of legal interpretation, enhances the transparency of normative evolution, and provides a foundation for more evidence-based interaction among legislators, researchers, and civil society. Recent works employing large language models for empirical legal studies demonstrate the growing potential of AI-assisted thematic and linguistic analysis within judicial and legislative corpora [14]. Frameworks developed for qualitative content analysis, such as the Leipzig Corpus Miner, further contextualize the significance of connecting linguistic analysis with institutional and semantic interpretation [16].

Materials and methods

Data

The dataset used in this research was collected from the official legal information portal adilet.zan.kz, which hosts the digital versions of Kazakhstan's legislative acts. To enable large-scale and reproducible data collection, a Python-based Scrapy parser was developed. This parser can automatically retrieve the full text of any legal code from adilet.zan.kz using its unique document identifier (e.g.,K1400000235).

During extraction, all textual and structural elements – including sections, chapters, articles, and paragraphs – are normalized and serialized into a consistent, machine-readable JSON schema. The framework is designed to support automated dataset updates: when amendments or new versions of a Code are published, the parser can re-collect and synchronize the data without manual intervention. This functionality allows the framework to be extended to future analyses of other Kazakhstani legal codes (e.g., Civil, Criminal, or Tax Codes), ensuring both scalability and long-term maintainability of the corpus.

The parser was applied to the Code of Administrative Offenses of the Republic of Kazakhstan (K1400000235), performing full hierarchical segmentation into:

- Sections (Разделы),
- Chapters (Главы),
- Articles (Статьи),
- Paragraphs (Пункты).

Each paragraph is stored with metadata describing its position in the hierarchy and textual content. The resulting dataset contains 3,336 paragraph-level units, represented through the following metadata fields:

- doc_id – unique document identifier
- lang – language of the text
- url – source link on Adilet.zan.kz
- title – official title of the Code
- status – legal status (e.g., updated)
- citation – formal legal citation
- section – section title
- chapter – chapter title
- article_id – internal article ID
- article_title – article heading
- article_text – full article text
- paragraphs – list of paragraph-level units
- notes – amendment and commentary notes
- links – cross-references to other laws

This structured representation was subsequently exported in JSON format, ensuring hierarchical consistency and facilitating downstream linguistic and statistical analysis. A descriptive statistical overview of the resulting corpus – including the number of articles, paragraphs, and amendment notes – is presented in Section 4.1.

Preprocessing

All paragraph-level texts were preprocessed using the spaCy natural language processing library with the ru_core_news_sm language model. This choice was motivated by the need for a deterministic, transparent, and linguistically grounded preprocessing pipeline suitable for large-scale legal corpora.

Although large language models (LLMs) offer advanced semantic interpretation, their outputs are often non-deterministic and depend on model-specific context, prompts, or sampling parameters. Moreover, LLMs require substantial computational resources, making them less suitable for foundational linguistic preprocessing tasks such as tokenization and lemmatization, where consistency, efficiency, and reproducibility are essential.

The spaCy framework ensures robust tokenization, lemmatization, and part-of-speech tagging for Russian text, while operating fully offline without reliance on external APIs or network connectivity. The lightweight ru_core_news_sm model was specifically chosen to balance computational efficiency and linguistic accuracy, enabling rapid processing of thousands of legal paragraphs while preserving the precision required for lexical frequency and morphological analysis.

The text preprocessing pipeline comprised the following sequential stages:

1. Text normalization – conversion of all text to lowercase and removal of punctuation, numerical symbols, and extraneous whitespace to ensure uniform lexical representation.
2. Tokenization – segmentation of normalized text into individual lexical units (tokens).
3. Stop-word filtering – exclusion of high-frequency Russian functional words using the stop_words_ru corpus to eliminate non-informative linguistic elements.
4. Lemmatization – reduction of each token to its canonical (dictionary) form to consolidate morphological variants of the same lexical item.
5. Storage of results - creation of a new field, tokens, containing the final list of lemmatized words for each paragraph.

This standardized preprocessing pipeline ensured linguistic consistency and reproducibility across the entire dataset, forming the basis for subsequent lexical and statistical analyses.

Each paragraph was represented in a normalized JSON structure that preserves both its hierarchical and linguistic attributes. An example entry is shown below:

{
“doc_id”: “K1400000235”,
“section”: “РАЗДЕЛ 1. ОБЩИЕ ПОЛОЖЕНИЯ”,
“chapter”: “Глава 1. ЗАКОНОДАТЕЛЬСТВО ОБ АДМИНИСТРАТИВНЫХ ПРАВОНАРУШЕНИЯХ”,
“article_id”: “z4”,
“paragraph_id”: “z5”,
“text”: “Законодательство Республики Казахстан об административных правонарушениях состоит из настоящего кодекса.”,
“tokens”: [“законодательство”, “республика”, “казахстан”, “административный”, “правонарушение”, “кодекс”]
}

This structured representation ensures direct correspondence between the legal text and its computationally processed form, facilitating precise alignment of linguistic and legal features across the corpus. The resulting tokenized corpus was used as the primary input for lexical frequency analysis, keyword distribution studies, and visualization through frequency plots and thematic heatmaps.

Furthermore, the framework was designed to remain extensible and interoperable with advanced NLP pipelines. In particular, it supports potential integration with large language models (LLMs) for semantic enrichment tasks such as contextual classification, topic expansion, and automated summarization. These capabilities, however, were not utilized in the present study, which focuses on deterministic linguistic and statistical analyses.

Processing

The processing stage includes the computational procedures applied after the initial corpus collection and preprocessing. This stage describes the extraction of thematic features, the selection and grouping of legally relevant keywords, and the application of quantitative analytical methods to the normalized text. The objective of this module is to transform the lemmatized corpus into a structured representation suitable for lexical frequency analysis, semantic field exploration.

Subsections 3.3.1 and 3.3.2 describe the construction of the legal keyword lexicon used for thematic analysis and the computational tools applied throughout the study. These components form the core of the quantitative workflow used to examine the lexical structure of the CAO RK.

In addition, Subsection 3.3.3 outlines the hyperparameters used in the numerical experiments, ensuring transparency and reproducibility of the analytical results.

Keyword Selection

To enable thematic and semantic analysis, a specialized lexicon of key legal concepts was constructed. The lexicon reflects the principal conceptual domains of administrative law and was manually curated based on the terminology and recurrent expressions found in the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK).

Each category corresponds to a semantic field encompassing morphologically related words and stems. This approach allows for the inclusion of lexical variation (e.g., inflectional and derivational forms) while maintaining semantic coherence across categories. The principal thematic categories and representative keyword stems are listed below:

- Sanctions – штраф, взыскан, санкц, наказан
- Responsibility – ответственн, виновн
- Offense – правонарушен, нарушен
- Person – лицо, гражданин, должностн
- Entrepreneurship – предприниматель, бизнес, юрлиц
- Authority – орган, комитет, инспекц, министерств
- Court – суд, судья, кассац, апелляц
- Property – имущество, собственност, доход
- Size / Measure – размер, расчетн, месячн, МРП
- Case – дело, производств, рассмотрен
- State – казахстан, республика, кодекс

Occurrences of these keyword stems were tracked across paragraphs, articles, and chapters to quantify their frequency, relative prominence, and contextual distribution within the Code. Grouping tokens by morphological stems (e.g., ответственн → ответственность, ответственный) ensured robust lexical matching and prevented over-fragmentation of semantically equivalent terms.

This curated keyword lexicon thus served as the analytical foundation for quantitative thematic profiling and semantic field visualization in subsequent stages of analysis.

Tools

All stages of the analysis were conducted in a Python 3.11 environment using Jupyter Notebook as the interactive development platform. The computational workflow relied exclusively on open-source libraries widely adopted in the fields of data analysis and natural language processing. The primary libraries and their respective functions are summarized below:

- pandas – data ingestion, transformation, and aggregation;
- spaCy – tokenization and lemmatization of Russian texts (ru_core_news_sm model);
- NLTK – stop-word filtering and frequency-based text statistics;
- matplotlib and seaborn – visualization of lexical distributions and thematic heatmaps;
- pathlib and json – structured I/O and corpus serialization.

All scripts were executed within a notebook-based reproducible workflow, ensuring full transparency, modularity, and consistency across the stages of data preprocessing, keyword extraction, and statistical analysis.

Hyperparameters for numerical experiments

Latent Dirichlet Allocation (LDA) was used to examine higher-level thematic structures within the corpus. Input matrix was constructed with a bag-of-words representation using CountVectorizer. To ensure transparency and reproducibility, all hyperparameters and configuration settings are reported below.

Vectorization settings (CountVectorizer):

- max_df = 0.9 – terms appearing in more than 90% of paragraphs were excluded as overly frequent;
- min_df = 3 – terms occurring in fewer than three paragraphs were removed to reduce sparsity;
- stop words – no additional stop-word lists were applied beyond earlier preprocessing steps;
- tokenization – pre-tokenized and lemmatized tokens were joined into whitespace-separated strings.

LDA model settings (LatentDirichletAllocation):

- n_components = 5 – the number of latent topics;
- random_state = 42 – fixed seed for reproducibility;
- learning_method = ‘batch’;
- max_iter = 10;
- doc_topic_prior (α) = 1 / n_components;
- topic_word_prior (β) = 1 / n_components.

These settings were chosen to balance topic interpretability and computational efficiency, given the relatively compact size of the corpus. The resulting model outputs were transformed into interactive visualizations using pyLDAvis.

The visualization module was supplied with:

- topic-term distributions – (lda.components_ normalized row-wise),
- document-topic distributions – (lda.transform(X)),
- document lengths – computed from token counts,
- vocabulary – extracted from the vectorizer,
- term frequencies – aggregated across all documents.

This configuration enabled an interpretable inspection of topic coherence, dominant terms, and the overall thematic landscape of the Code.

Results

Descriptive Overview of the CAO RK Corpus

Prior to conducting lexical and thematic analyses, the corpus was examined descriptively to establish its structural composition and quantitative characteristics. Table 1 presents the summary statistics of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK) corpus compiled for this study.

The dataset comprises 1,050 articles and 3,336 paragraphs, with an average paragraph length of approximately 40 words. The longest paragraph contains more than 2,000 words, reflecting the extensive procedural detail typical of certain sections of administrative law. In total, 781 legislative notes (сноски) were identified, documenting amendments introduced between 2014 and 2025. These amendment annotations provide clear evidence of the Code's continuous normative evolution over the past decade.

Table 1 – Summary characteristics of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

Metric	Value
Total number of articles (статьи)	1,050
Total number of paragraphs (пункты)	3,336
Average paragraph length (words)	39.6
Maximum paragraph length (words)	2,042
Number of amendment notes (сноски)	781
Year range of amendments	2014–2025

Word Frequency Distribution

The twenty most frequent words in the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK) are predominantly associated with quantitative measures (e.g., размер, показатель, месячный расчётный показатель) and legal actors (e.g., лицо, субъект).

This lexical pattern highlights the Code's focus on defining responsibilities, entities, and the scale of administrative penalties – central dimensions of administrative law.

Figure 1, presented directly below, presents a horizontal bar chart depicting the relative frequencies of the most common lexical units in the corpus after lemmatization and stop-word removal. The clear predominance of размер ("amount") and лицо ("person") underscores the quantitative and person-centric orientation of the CAO RK, revealing its emphasis on measurable sanctions and the identification of liable subjects.

Distribution of Key Terms Across Chapters

To investigate the thematic organization of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), a set of keywords was grouped into semantic categories representing the principal conceptual domains of administrative law. These categories encompass notions of sanctions, responsibility, administrative subjects, institutional governance, economic relations, and procedural operations.

Each paragraph of the Code was tokenized and normalized into lemmas, after which the occurrences of keyword stems were aggregated at the chapter level. This procedure enabled a comparative assessment of the relative prominence of distinct legal domains within the legislative corpus.

The results demonstrate a non-uniform lexical distribution across chapters. Chapters addressing personal rights, entrepreneurial activity, and public administration exhibit a particularly high density of sanction-related and institutional vocabulary (e.g., санкции, ответственность, штраф), indicating an emphasis on punitive and regulatory enforcement mechanisms. By contrast, chapters governing

procedural processes show elevated frequencies of lexemes associated with дело (“case”) and орган (“authority”), reflecting the procedural dimension of administrative adjudication.

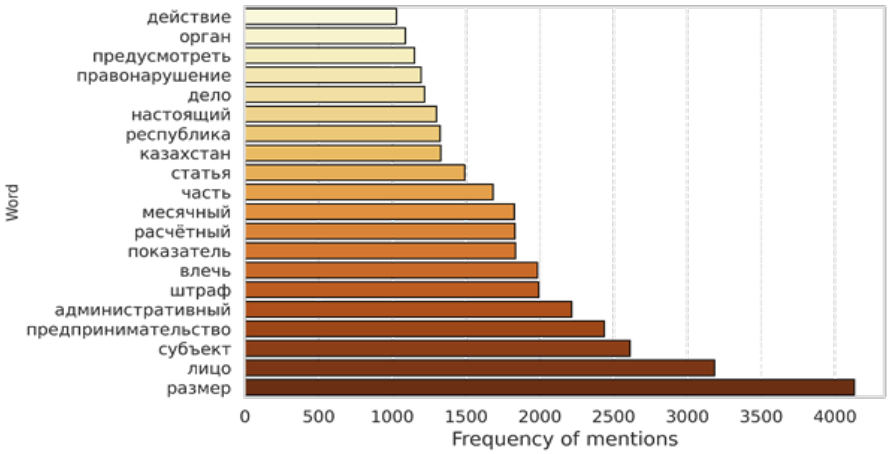

Figure 1 – Top-20 most frequent words in the Code of Administrative Offenses

Overall, the analysis suggests that the CAO RK is lexically organized around several semantic clusters corresponding to its functional domains–sanction-related, procedural, administrative, and economic. This thematic stratification highlights the internal diversity of administrative law, where the distribution of linguistic features reflects the differentiated structure of legal regulation.

Figure 2, presented immediately after this paragraph, visualizes the ten chapters with the highest cumulative frequency of thematic keywords. The chart demonstrates clear disparities in the lexical salience of legal concepts across chapters, reflecting the internal hierarchy of regulatory focus areas within the Code.

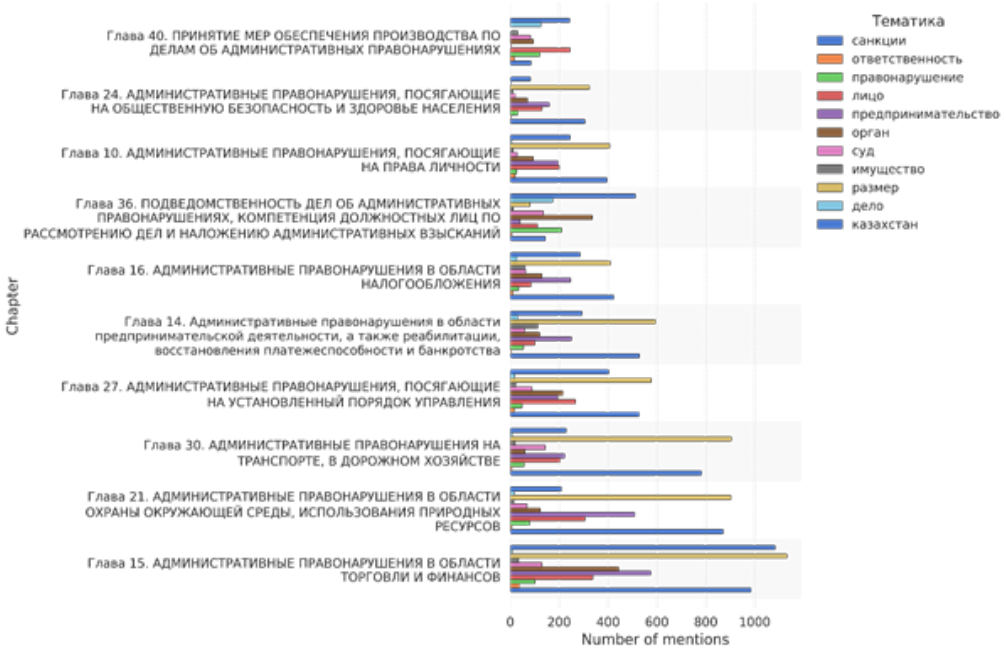

Figure 2 – Top-10 chapters of the CAO RK by frequency of thematic keywords (grouped by semantic categories)

Thematic Heatmap

A thematic heatmap (Figure 3) was constructed to visualize the distribution of key legal concepts across the structural sections of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK). The visualization reveals distinct lexical clustering patterns aligned with the primary domains of administrative regulation.

The semantic fields associated with санкции (sanctions), ответственность (responsibility), and лицо (person) exhibit consistently high frequencies, underscoring their central role in the conceptual architecture of the Code. Conversely, terms connected to имущество (property) and суд (court) are concentrated in procedural and enforcement-related sections, indicating specialized legal subdomains.

Sections regulating entrepreneurship and property display elevated lexical density across multiple thematic categories – particularly in sanction-related vocabulary – reflecting the Code's emphasis on economic accountability and regulatory oversight.

Collectively, the heatmap demonstrates that the CAO RK is not lexically uniform but instead organized around discrete thematic cores corresponding to its institutional and functional structure.

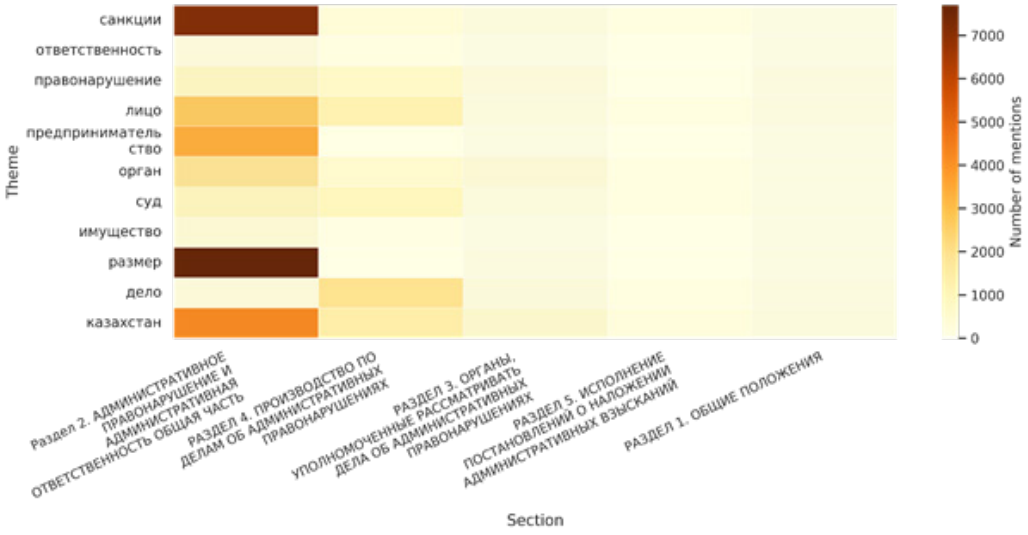

Figure 3 – Thematic distribution of legal categories across sections
of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

Dominant Chapter–Term Intersection

To highlight the primary lexical orientation of the most content-dense chapters, a "chapter × term" intersection analysis was conducted.

For each chapter, the thematic category with the highest frequency was identified as its dominant lexical field. Among the ten chapters exhibiting the greatest overall lexical density, the field размер ("measure" or "amount") emerges as the most recurrent leading term, especially in sections regulating trade, finance, and administrative enforcement. This prevalence reflects the quantitative character of sanction-related provisions, where penalties are systematically defined through monetary or procedural measures.

Other dominant categories include санкции (sanctions), лицо (person), and казахстан, the latter appearing primarily in chapters devoted to procedural jurisdiction and institutional authority.

The pattern illustrates that the CAO RK emphasizes proportionality and standardization of sanctions, embedding numerical evaluation directly into the language of legal accountability (Figure 4).

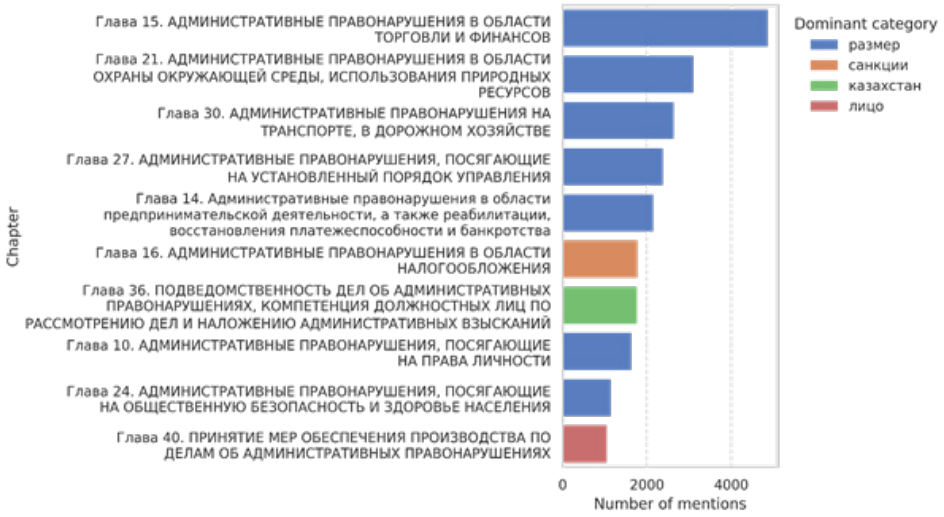

Figure 4 – Dominant thematic category in the ten most content-dense chapters of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

Article Length Distribution

The dual-panel histogram (Figure 5) provides a detailed view of the lexical structure of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK).

The left panel displays the overall distribution of article lengths, illustrating the extreme right-skew typical of codified legal texts.

The right panel zooms into the 0–500 word range, revealing that the majority of articles are compact, generally containing fewer than 250–300 words.

This pattern indicates that while most provisions are concise and normatively focused, a limited number of articles – primarily those describing procedural, sanctioning, or jurisdictional rules – expand substantially in length to accommodate complex legal formulations.

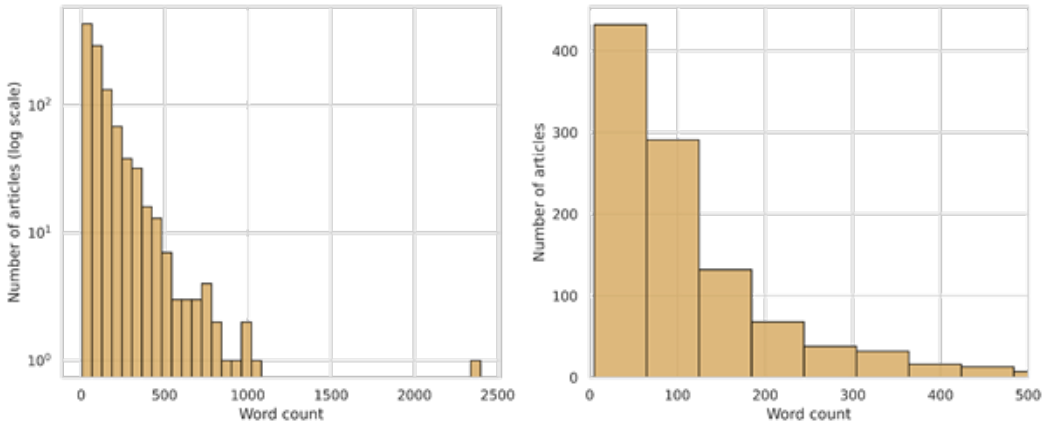

Figure 5 – Distribution of article lengths (in words) in the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

The left panel (logarithmic scale) shows the overall right-skewed distribution, while the right panel highlights the concentration of shorter articles below 500 words.

Amendment Dynamics

The temporal distribution of legislative amendments (Figure 6) highlights distinct cycles of revision within the Code of Administrative Offenses of the Republic of Kazakhstan.

Amendment metadata were extracted from article-level footnotes (сноски), capturing formal references to legislative acts and their years of adoption.

The histogram reveals a major legislative peak in 2017, marking the most extensive wave of modifications during the observed period. This surge corresponds to a comprehensive administrative reform aimed at refining sanctioning procedures and institutional responsibilities across multiple chapters of the Code. Moderate amendment activity is also observed in 2014–2015, reflecting the transitional phase following the enactment of the CAO RK, when the legal system was adjusting to the new codified framework. Subsequent years (2018–2022) display more selective revision patterns, including targeted changes to procedural and jurisdictional provisions.

Although the COVID-19 pandemic (2020–2021) did not produce a large quantitative spike in amendments, the updates introduced during this period often addressed public health enforcement, mobility regulation, and emergency governance, signaling a functional adaptation of administrative law to crisis conditions.

The mild uptick in 2025 suggests the beginning of another adjustment cycle aligned with ongoing modernization efforts.

Overall, the amendment dynamics illustrate a nonlinear but adaptive evolution of Kazakhstan's administrative legislation, alternating between large-scale reform phases and incremental fine-tuning periods.

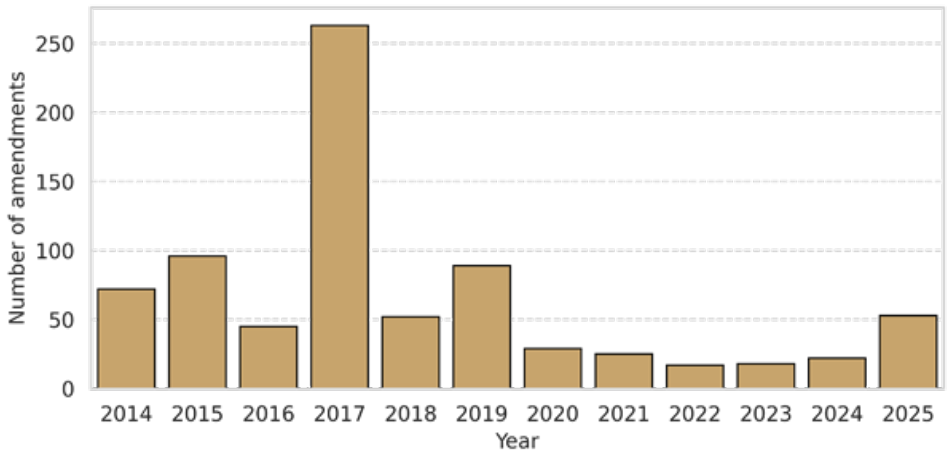

Figure 6 – Distribution of amendment years across articles
of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

Topic Modeling of Legal Texts

To uncover latent semantic structures within the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), a topic modeling approach was applied using the Latent Dirichlet Allocation (LDA) algorithm. This probabilistic model identifies recurring lexical patterns by representing each article as a mixture of topics and each topic as a distribution over words. The goal was to capture underlying conceptual domains of administrative regulation that may not be explicitly defined by the Code's formal hierarchy.

The model was trained with five latent topics (n = 5), selected empirically to optimize semantic coherence and interpretability. The ten most probable terms for each topic, together with their domain-specific interpretations, are presented in Table 2. These representative term clusters served as the basis for qualitative labeling of topics according to their legal and contextual significance.

The five topics collectively reflect the dual nature of the CAO RK's discourse – a balance between quantitative sanctioning logic (Topics 1, 2, and 4) and institutional-procedural governance (Topics 3 and 5).

Table 2 – Topical structure of the CAO RK identified by LDA (Five-Topic Model)

Topic	Representative Terms	Interpretation
1	месячный, расчётный, показатель, штраф, влечь, организация, размер, лицо, республика, казахстан	Quantitative and procedural norms, focusing on the formal specification of sanctions and fiscal measures.
2	крупный, влечь, штраф, предпринимательство, субъект, размер	Economic and entrepreneurial regulation, emphasizing business responsibility and liability.
3	правонарушение, закон, действие, административный, орган, статья, часть	General principles and institutional structure of administrative law, describing legal actions and authorities.
4	расчётный, месячный, предусмотреть, статья, настоящий, штраф	Legislative and prescriptive formulations typical of codified sanctioning clauses.
5	кодекс, суд, должностной, орган, постановление, производство, правонарушение, дело	Judicial and procedural administration, covering courts, officials, and adjudicatory processes.

This thematic composition suggests that the Code's linguistic architecture is shaped by a systematic concern with measurement, enforcement, and procedural formality, rather than narrative or interpretive exposition.

In other words, the CAO RK encodes legal responsibility through a lexically rigid framework of proportionality and administrative control.

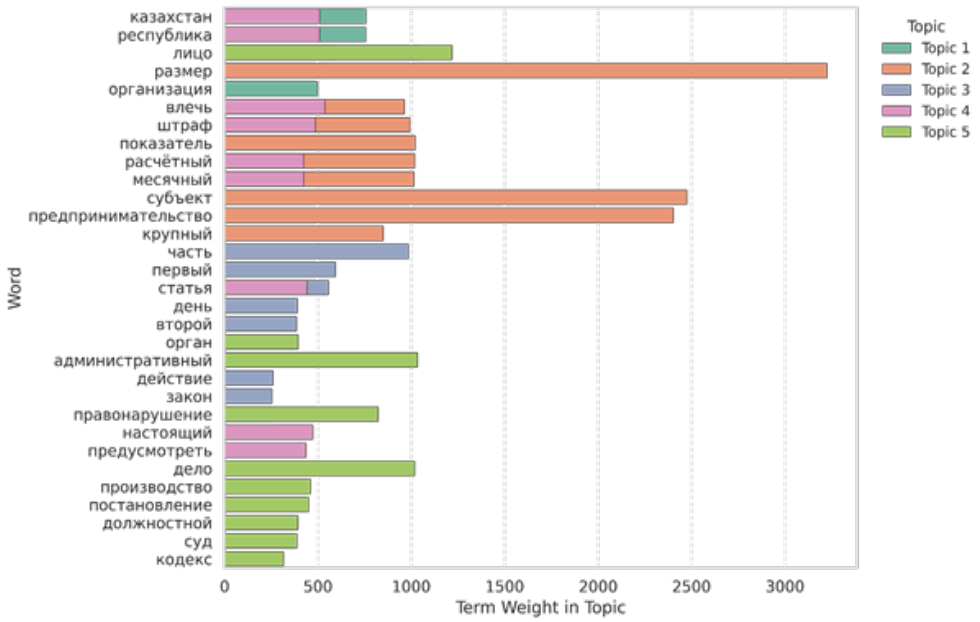

Figure 7 – Distribution of key terms across topics identified by Latent Dirichlet Allocation (LDA) in the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK)

Interactive Topic Visualization

To further explore the semantic topology of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), an interactive intertopic distance map was generated using pyLDAvis (Figure 8). This visualization projects the latent topics produced by the LDA model into a two-

dimensional semantic space via multidimensional scaling, allowing an intuitive assessment of topic proximity, distinctiveness, and lexical dominance.

The left panel displays the spatial arrangement of the five inferred topics, where the size of each circle represents its marginal proportion in the corpus.

The relatively large and well-separated clusters indicate that the LDA model captured thematically coherent and linguistically distinct domains of administrative regulation.

Topic 2 (highlighted in red) occupies the greatest share–approximately 29% of the total token space – and centers on vocabulary related to предпринимательство (“entrepreneurship”), субъект (“subject”), and размер (“measure”), corresponding to the economic and sanction-quantification dimension of the Code.

The right panel ranks the top-30 most relevant terms for the selected topic according to the relevance metric $\lambda = 1.0$, balancing frequency and exclusivity.

The prominence of terms such as предпринимательство, штраф, and организация confirms that the model effectively isolates the language of regulatory enforcement and business accountability.

Taken together, the intertopic map reveals a structured semantic architecture, where distinct lexical clusters correspond to sanctioning mechanisms, procedural norms, institutional roles, and economic regulation.

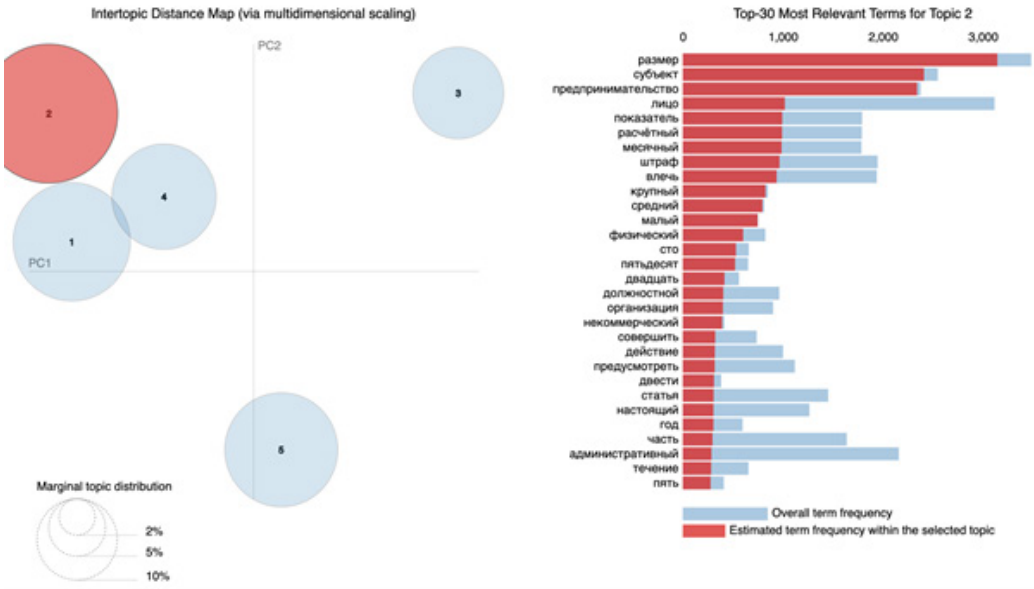

Figure 8 – Intertopic distance map and term relevance visualization
for the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK),
generated via Latent Dirichlet Allocation (LDA) and pyLDAvis

Discussion

The linguistic analysis of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK) reveals that its textual architecture distinctly embodies the normative-punitive nature of administrative law.

The overwhelming prevalence of vocabulary related to sanctions (санкции) and responsibility (ответственность) indicates that the Code primarily operates as an instrument for the definition, categorization, and enforcement of administrative penalties.

At the same time, the frequent occurrence of terms such as лицо (“person”) and субъект (“subject”) highlights the Code’s focus on the individualization of legal liability, consistent with the modern principle of personal accountability in administrative justice.

In contrast, the relative scarcity of judicial lexemes – such as суд (“court”) and жалоба (“appeal”) – underscores the subsidiary role of judicial institutions, where administrative enforcement and executive oversight remain dominant.

From a broader linguistic perspective, this quantitative profile confirms that the CAO RK is not merely a collection of prohibitive norms but a codified linguistic framework through which state authority is operationalized via structured legal terminology.

The findings therefore complement traditional doctrinal interpretations by showing how lexical hierarchies mirror institutional hierarchies within Kazakhstan’s system of administrative governance.

Recent advances in Kazakh computational linguistics reinforce the relevance of this approach. The development of large-scale instruction-tuning datasets for government and legal domains, together with corpora such as KazQAD and the Aligned Kazakh–Russian Parallel Corpus, demonstrates growing national capacity for data-driven linguistic analysis [6–8]. However, these initiatives primarily address general or bilingual text processing, whereas the present study focuses specifically on formal legislative discourse, providing a complementary structural and thematic dimension to the evolving landscape of Kazakh-language NLP. This alignment between legal linguistics and computational modeling marks an important step toward creating comprehensive resources for automated legal analytics and low-resource language modeling in Kazakhstan.

Conclusion

This study employed quantitative linguistic analysis to examine the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), uncovering both structural and thematic regularities within its legal discourse. The findings reveal a marked predominance of sanction- and responsibility-related vocabulary, underscoring the punitive and regulatory orientation of the Code and its function as a linguistic instrument of state authority.

Beyond descriptive insights, the research demonstrates the applicability of computational text analysis to the study of legal language – offering a scalable, data-driven framework for comparative legal linguistics, corpus-based jurisprudence, and the modernization of legislative analytics in Kazakhstan.

These results complement emerging initiatives in Kazakh natural language processing, such as the development of instruction-tuned datasets for legal and governmental text, the KazQAD question-answering corpus, and the Kazakh–Russian parallel legal corpus [6–8].

Future research will extend this approach to other codified acts–such as the Civil, Criminal, and Tax Codes – to explore differences in lexical density, thematic diversity, and cross-referential complexity, thereby advancing a data-driven understanding of national legal language systems.

In this broader context, the present work contributes a structured analytical pipeline and an annotated legislative corpus that can support future efforts toward Kazakh legal language modeling and semantic text alignment across legal codes.

Looking ahead, the framework developed here remains fully compatible with LLM-based semantic enrichment – including contextual classification, summarization, and translation consistency evaluation.

Future research will extend this approach to other codified acts – such as the Civil, Criminal, and Tax Codes – to explore differences in lexical density, thematic diversity, and cross-referential complexity, thereby advancing a data-driven understanding of national legal language systems and their evolution.

Limitations and Future Directions

While the present analysis provides a comprehensive quantitative overview of the Code of Administrative Offenses of the Republic of Kazakhstan (CAO RK), several methodological and interpretive limitations should be acknowledged. First, the study is confined to a single legal corpus, which constrains cross-institutional generalization. Extending the approach to other Kazakhstani codes and to comparative datasets from neighboring jurisdictions would enable a broader understanding of regional legal-linguistic patterns.

Second, the analysis relies primarily on surface-level lexical statistics and topic modeling, which, although informative, do not fully capture syntactic dependencies, modal constructions, or intertextual citation structures that play a central role in legal semantics. Incorporating dependency parsing, named-entity recognition, and semantic role labeling in future work would enhance interpretive depth.

Third, while the LDA-based topic modeling revealed coherent thematic clusters, it remains probabilistic rather than interpretive; integrating these findings with expert legal annotation could bridge the gap between computational and doctrinal perspectives.

Future research should also explore temporal evolution by tracking linguistic change across successive amendments, and should consider applying embedding-based language models (e.g., BERT or RoBERTa trained on legal corpora) to quantify semantic shifts within Kazakhstan's evolving legal framework.

REFERENCES

1 Theory and Methodology of the World's National Linguistic Corpora. Linguistics Journal of Eurasia, 14(3), 33–45 (2022).

2 Tokatov, R.A., Akimzhanova, M.T. On the accuracy of the texts of the Civil Code of the Republic of Kazakhstan (General Part) in the Kazakh and Russian languages. Bulletin of L.N. Gumilyov Eurasian National University. Law Series, 3, 135–141 (2021). https://doi.org/10.31489/2021l3/135-141.

3 Ilyassova, G.A. Problems of ensuring authenticity of texts in Kazakh and Russian in the Civil Procedure Code of the Republic of Kazakhstan. Bulletin of L.N. Gumilyov Eurasian National University. Law Series, 3, 71–78 (2022). https://doi.org/10.31489/2022l3/71-78.

4 Ilyassova, G.A. Issues of application of terms in the state language in civil legislation (according to the text of the special part of the Civil Code of the Republic of Kazakhstan). Vestnik Akademii Upravleniya, 69(2), 123–134 (2023). https://doi.org/10.47649/vau.2023.v69.i2.14.

5 Zhanzhigitov, S.Zh. Linguistic strategies of legal communication in digital environments: The case of the PravMedia online forum. Bulletin of L.N. Gumilyov Eurasian National University. Philology Series, 3, 41–49 (2024). https://doi.org/10.31489/2024ph3/41-49.

6 Yeshpanov, R., Efimov, P., Boytsov, L., Shalkarbayuli, A., Braslavski, P. KazQAD: Kazakh open-domain question answering dataset. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024). Torino, Italy: European Language Resources Association (ELRA), 2024, pp. 9645–9656.

7 Khairova, N., Kolesnyk, A., Mamyrbayev, O., Mukhsina, K. The Aligned Kazakh–Russian Parallel Corpus Focused on the Criminal Theme. Proceedings of the International Conference on Computational Linguistics and Intelligent Systems (COLINS), 2019, pp. 116–125.

8 Baisalov, A., Kenzhegulov, Y., Alimzhanova, Z. Instruction tuning on public government and cultural data for low-resource language: A case study in Kazakh. Proceedings of the 2024 Conference on Computational Linguistics for Low-Resource Languages, 2024 (Preprint available on arXiv).

9 Formation of the State Language as the Language of the Law. Bulletin of Law and State, 2, 56–68 (2022).

10 Kolesnik, A., Khairova, N. Use of linguistic criteria for estimating the quality of Wikipedia articles. Proceedings of the 1st International Conference on Computational Linguistics and Intelligent Systems (COLINS), 2017, pp. 207–215.

11 Khairova, N., Mamyrbayev, O., Rizun, N., Razno, M., Ybytayeva, G. A parallel corpus-based approach to crime event extraction for low-resource languages. IEEE Access, 11, 54093–54111 (2023).

12 Yeshpanov, R., Varol, H.A. KazSAnDRA: Kazakh sentiment analysis dataset of reviews and attitudes. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024), Torino, Italy: European Language Resources Association (ELRA), 2024.

13 Karipova, A., Serikbekova, S., Aralbekov, G., Tuleugaliyeva, Zh., Sarsenova, A. Comparative analysis of administrative liability for driving while intoxicated in the Commonwealth of Independent States. Hrvatska i komparativna javna uprava. Croatian and Comparative Public Administration, 24(4), 889–910 (2024).

14 Drápal, J., Westermann, H., Savelka, J. Using large language models to support thematic analysis in empirical legal studies. Proceedings of the Thirty-sixth Annual Conference on Legal Knowledge and Information Systems (JURIX 2023). Maastricht, The Netherlands: IOS Press, 2023, pp. 65–74.

15 Malik, V., Sanjay, R., Guha, S.K., Hazarika, A., Nigam, S.K., Bhattacharya A., Modi A. Semantic segmentation of legal documents via rhetorical roles. Proceedings of the Natural Legal Language Processing Workshop (NLLP 2022). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022, pp. 132–142.

16 Niekler, A., Wiedemann, G., Heyer, G. Leipzig Corpus Miner: A text mining infrastructure for qualitative data analysis. Proceedings of the Terminology and Knowledge Engineering Conference (TKE 2014). Berlin, Germany, 2014, pp. 441–450.

[1]*Мухсимбаев Б.,
докторант, ORCID ID: 0009-0008-4606-3628,
*e-mail: b.mukhsimbaev@kbtu.kz
[1]Пак А.,
PhD, профессор, ORCID ID: 0000-0002-8685-9355,
e-mail: a.pak@kbtu.kz
[1]Куралбаев А.,
докторант, ORCID ID: 0009-0001-0811-5385,
e-mail: a.kuralbaev@kbtu.kz

[1]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ӘКІМШІЛІК ҚҰҚЫҚ БҰЗУШЫЛЫҚ ТУРАЛЫ КОДЕКСІН ЛЕКСИКАЛЫҚ ЖӘНЕ ТАҚЫРЫПТЫҚ ТАЛДАУҒА АРНАЛҒАН ЕСЕПТЕУ КОНВЕЙЕРІ

Аңдатпа

Бұл зерттеу Қазақстан Республикасының Әкімшілік құқық бұзушылықтар туралы кодексі (ӘҚБтК, К1400000235) мәтінінің автоматтандырылған лингвистикалық және құрылымдық талдауына арналған есептеуіш талдау жолын ұсынады. Ұсынылған жұмыс процесі деректерді жинау, мәтінді алдын ала өңдеу, токенизация, кілт сөздерді анықтау, семантикалық топтастыру және визуализацияны қамтиды. Бұл кезеңдер Python тіліндегі табиғи тілдерді өңдеу (NLP) және статистикалық әдістерді біріктіреді. Ұсынылған жүйе лексикалық, тақырыптық және сандық лингвистикалық талдауларды бірізді тізбекке біріктіріп, Кодекстің иерархиялық құрылымы (бөлімдер, тараулар және баптар) бойынша жиілік үлестірімдерін, семантикалық өрістер мен жасырын тақырыптарды анықтауға мүмкіндік береді. ӘҚБтК корпусын талдау бірнеше ерекше тілдік заңдылықтарды анықтады: санкциялар мен жауапкершілікке қатысты сөздердің (айыппұл, жауапкершілік, құқық бұзушылық) басым болуы, экономикалық және рәсімдік құқық бұзушылықтарға арналған тарауларда жоғары лексикалық тығыздықтың байқалуы, сондай-ақ әкімшілік құқықтың нормативтік-жазалаушылық сипатын бейнелейтін тақырыптық шоғырланулар. Жиілік гистограммалары, тақырыптық жылу карталары және тақырыптық карталар сияқты визуализация әдістері заң мәтіндерін сандық тұрғыдан зерттеудің әлеуетін көрсетеді. Жалпы алғанда, ұсынылған әдістеме салыстырмалы құқықтық лингвистика, заңнаманы автоматты түрде мониторингілеу және Қазақстандағы құқықтық аналитиканы жаңғырту үшін ауқымды негіз қалайды.

Тірек сөздер: әкімшілік құқық, заң мәтінін талдау, табиғи тілді өңдеу, есептеуіш құқықтық лингвистика, жиілік талдауы, тақырыптық модельдеу, құқықтық информатика.

[1]Мухсимбаев Б.,
докторант, ORCID ID: 0009-0008-4606-3628,
e-mail: b.mukhsimbaev@kbtu.kz
[1]Пак А.,
PhD, профессор, ORCID ID: 0000-0002-8685-9355,
e-mail: a.pak@kbtu.kz
[1]Куралбаев А.,
докторант, ORCID ID: 0009-0001-0811-5385,
e-mail: a.kuralbaev@kbtu.kz

[1]Казахстанско-Британский технический университет, г. Алматы, Казахстан

ВЫЧИСЛИТЕЛЬНЫЙ КОНВЕЙЕР ДЛЯ ЛЕКСИЧЕСКОГО И ТЕМАТИЧЕСКОГО АНАЛИЗА КОДЕКСА ОБ АДМИНИСТРАТИВНЫХ ПРАВОНАРУШЕНИЯХ РЕСПУБЛИКИ КАЗАХСТАН

Аннотация

В статье представлен вычислительный конвейер для автоматизированного лингвистического и структурного анализа юридических текстов на примере Кодекса Республики Казахстан об административных правонарушениях (КоАП РК, K1400000235). Предложенный рабочий процесс объединяет этапы сбора данных, предобработки текста, токенизации, извлечения ключевых слов, семантической кластеризации и визуализации с применением методов обработки естественного языка (NLP) и статистического анализа на Python. Разработанный конвейер сочетает лексический, тематический и количественный лингвистический анализ в единую последовательную систему, что позволяет выявлять частотные распределения, семантические поля и скрытые темы в иерархической структуре Кодекса (разделы, главы, статьи). Анализ корпуса КоАП РК выявил ряд характерных языковых закономерностей: преобладание лексики, связанной с санкциями и ответственностью (штраф, ответственность, правонарушение), повышенную лексическую плотность в главах, регулирующих экономические и процессуальные правонарушения, а также тематические кластеры, отражающие нормативно-карательную направленность административного права. Визуализационные методы, такие как частотные гистограммы, тематические тепловые карты и топик-карты, демонстрируют потенциал конвейера для количественного исследования законодательного языка. В целом представленная методология формирует масштабируемую основу для сравнительной юридической лингвистики, автоматизированного мониторинга законодательства и модернизации правовой аналитики в Казахстане.

Ключевые слова: административное право, анализ юридических текстов, обработка естественного языка, вычислительная юридическая лингвистика, частотный анализ, тематическое моделирование, правовая информатика.

Article submission date: 07.11.2025