Describing Images: Linguistic Signatures of AI and Humans

Tatiana A. Litvinova, Viktoriya A. Molchanova
Voronezh State Pedagogical University
Voronezh, Russia
centr_rus_yaz@mail.ru

Olga V. Dekhnich
Belgorod State University
Belgorod, Russia
dekhnich@bsuedu.ru

Pavel V. Seredin
Voronezh State University
Voronezh, Russia
paul@phys.vsu.ru

***Abstract*—The rapid growth of multimodal large language models (LLMs) call for a detailed analysis of the linguistic features in AI-generated, image-conditioned descriptions. Using a paired corpus of AI and human descriptions of artworks from the Hermitage collection, we developed a comprehensive authorship framework that incorporates measures of lexical diversity, morphological density and entropy, syntactic roles and complexity, and a novel set of semantic oppositions designed to infer modality engagement (sensory, cognitive, emotional, etc.). We built interpretable linear classifiers (elastic net logistic regression and ridge logistic regression) and tested them on 100 unseen images, each paired with one AI and one human caption. These classifiers achieved high balanced accuracy (0.9). AI-generated captions showed significantly higher levels of subordination, verbs per sentence, object-type syntactic relations, and longer dependency distances, along with greater engagement with cognition, size and intensity, time, socialness, and positive emotions, as captured by our semantic opposition features, as well as increased lexical diversity. In contrast, human captions demonstrated stronger engagement with somatic, visual, and motor concepts, and a higher frequency of adjectival modifiers that contribute to scene setting and attribute description. Beyond average differences, AI-generated texts were more homogeneous. Prompts that explicitly request concrete visual evidence along with limited subordination and reduced scalar intensity can guide outputs toward human-like, perceptually grounded descriptions. This approach is particularly valuable for applications requiring explainable, evidence-based reporting, such as medical imaging captions.**

I. Introduction

The advent of multimodal large language models (MLLMs) incorporating vision capabilities has revolutionized the task of generating textual descriptions for images, known as image captioning – a longstanding challenge in both the computer vision and natural language processing [1-3]. These models can generate lengthy, detailed descriptions based on visual input, surpassing the short captions typical of traditional approaches. Research has demonstrated that state-of-the-art models, such as GPT-4o, perform on par or even surpass human-level performance in pairwise caption battles paradigm, at least for images primarily focused on everyday life scenarios [3].

As new models generate long, detailed image descriptions, not only does their ability to provide precise depictions of image content becomes increasingly important. Equally critical is the study of their linguistic characteristics. Developing prompting strategies that account for the linguistic differences between AI- and human-generated image descriptions is essential. This especially true in safety-critical domains like medicine, where linguistic cues such as negation, uncertainty, hedging, and qualifiers can significantly alter clinical meaning (e.g., “no hemorrhage” versus “possible hemorrhage”). Feature-level analyses help ensure that models interpret and use these cues correctly. The use of overconfident language (e.g., “definite,” “clear”) when evidence is weak can be a risky. Linguistic profiling aids in calibrating the tone and claims of descriptions to align accurately with visual evidence. Interpretable features reveal how a model constructs its descriptions, enabling transparent analysis beyond relying solely on accuracy metrics [3], which is particularly important in medical image analysis. Furthermore, when images vary due to new scanners, lighting conditions, or styles, stable linguistic behaviors (such as concreteness and spatial terms) offer an additional measure of robustness beyond visual accuracy.

This study aims to systematically compare picture descriptions in Russian produced by humans and MLLM using a carefully curated dataset. We have developed a reproducible pipeline that integrates lexical, morphological, syntactic, and semantic features of both AI- and human-generated picture descriptions. A key aspect of our approach is a novel set of features we call neurobiology-motivated semantic oppositions, which offer quantitative contrasts along conceptual dimensions, such as abstract versus concrete, cognitive versus perceptual, evaluative versus descriptive, etc.

In addition to analyzing descriptive contrasts, we develop predictive models to determine whether these features can differentiate AI-generated captions from human-generated ones on unseen data (using “caption” as general term for all text paired with an image). We approach this as an authorship classification problem and train interpretable linear classifiers – specifically, elastic net logistic regression and ridge logistic regression – as well as, for comparison, a non-linear RBF-kernel SVM. To ensure robustness and interpretability, we control for text length using ANCOVA-style adjustments. This combination of descriptive and predictive analyses ensures that observed differences are both interpretable and generalizable.

The contributions of our work can be summarized as follows:

• a feature-centric benchmark for caption authorship: a replicable pipeline that uses interpretable linguistic signals to distinguish AI-generated picture descriptions from those written by humans;

• introduction of semantic oppositions: a novel, domain-independent set of features that capture a range of contrasts motivated by neurobiological research on word meaning;

• comprehensive analysis with practical implications: this integrated evaluation highlights the differences between AI and human-generated long image descriptions across multiple levels. These insights can be applied to prompting strategies to achieve the desired style. Such an approach is valuable in designing captioning systems that balance abstraction with perceptual grounding to produce explainable and clinically trustworthy outputs.

II. Related Work

The advancement of MLLMs has highlighted the challenge of analyzing and evaluating lengthy textual descriptions of images [2]. In recent years, numerous studies have compared AI-generated and human-generated image descriptions, particularly for medical images [4]. However, the primary focus has largely been on downstream task accuracy, such as object or lesion detection, abnormality characterization, and diagnostic output [5], or on general quality assessments typically based on metrics that measure similarity to human descriptions [2]. Another approach, proposed in a recent benchmark study [3], involved human raters evaluating both human- and AI-generated image descriptions using subjective criteria such as precision, informativeness, hallucination, and attention to detail. These metrics could be further improved by incorporating a broader range of linguistic measures beyond AI-human content similarity, provided such measures are available.

To the best of our knowledge, the linguistic properties of image textual descriptions have not been studied. Since MLLMs generate image descriptions using text generators, it is appropriate to frame linguistic analysis of such descriptions within the context of systematic differences between AI- and human-generated text.

As LLMs have grown more powerful and widely accessible, research has largely concentrated on their task performance rather than examining their writing style [6]. Nevertheless, several studies have identified systematic linguistic differences between AI-generated and human writing across various genres and model architectures [6-10]. These distinctions are especially pronounced in instruction-tuned models, which often display an informationally dense, noun-heavy style characterized by increased nominalizations, noun- and preposition-based phrases for abstract descriptions, fewer human subjects, reduced use of epistemic stance markers, and greater lexical diversity [6-10].

Human texts exhibit a broader range of sentence lengths, distinct patterns in dependency and constituent types, shorter constituents, and more optimized dependency distances [7-10]. Compared to text generated by LLMs, humans tend to express stronger negative emotions, such as fear and disgust, and less joy [10]. LLM outputs contain more numbers, symbols, and auxiliary words, indicating a more objective tone than human texts, along with a higher frequency of pronouns [9]. Additionally, human texts generally have simpler syntactic structures and more varied semantic content [9]. Newer models often resemble each other more closely and exhibit less variation than human texts, a phenomenon known as model collapse [6, 9]. Texts produced by ChatGPT excel in categories such as social processes, analytical style, cognition, attentional focus, and positive emotional tone, leading researchers to conclude that that LLMs can be "more human than human" [10].

The authors [6] highlight that their findings underscore the importance of incorporating linguistic structures to gain a deeper understanding of the capabilities and outputs of LLMs. They particularly emphasize the significant role of linguistic expertise and functional views of language in both the application and development of LLMs. At present, various linguistic perspectives are frequently overlooked during the development and internal evaluation of these models [6].

While very high classification accuracies are often reported in the literature – for example, random forests and lasso-penalized logistic regression reaching 93–98% in [6]—the problem of generalizability remains. In that study, models trained on one corpus frequently collapsed to chance-level performance when applied to a different corpus, underscoring that success in within-domain classification does not necessarily transfer across datasets or registers.

It is important to note that, in the studies mentioned above, surprisingly little attention has been given to controlling for the effect of text length. Most research has concentrated on English-language texts. Furthermore, to the best of our knowledge, this body of work does not include textual descriptions of images generated by MLLMs. Our study is the first to investigate this area.

III. Methods

A. Dataset

The dataset used in this study comprises textual descriptions of images from the Hermitage collections [11]. Each image is paired with one AI-generated caption and between three and twelve human-written captions. The AI captions were generated by a state-of-the-art MLLM developed by Yandex Research, which was in the process of being released at the time of the paper was submitted. According to the dataset creators (personal communication), the model was given images as input without any accompanying text examples. The prompts were provided in Russian, and the generated texts were also in Russian. The human segment of the dataset was produced by native Russian speakers who tasked with describing the same artworks used as inputs to the MLLM.

All captions are stored with metadata, including image_id, source (AI or HUMAN), caption_id, and text. Below is a sample AI-generated description (translated from Russian) of T. Gainsborough's "Portrait of a Lady in Blue".

AI description: *An eighteenth-century portrait in a realist manner shows a woman in a blue lace dress, wearing a pearl necklace and earrings. Her hair is styled high with a feather, and she faces the viewer with a slight smile. The dark background and directional lighting emphasize the fabric's texture and the intricate details of her hair.*

Human description: *In the painting, a young woman—approximately 30–35 years old—is portrayed with a high, voluminous ash-coloured wig topped by a feathered hat. She has brown eyes and delicate features. A black ribbon tied in a bow and a small cross pendant adorn her neck, and a watch is worn on her right wrist. She is dressed in white and light-blue fabrics that coordinate with her hat.*

We started with the complete collection of artworks (n = 902), each accompanied by one AI-generated description and between three and twelve human-generated descriptions. For this study, we limited the dataset to images that had all twelve human captions as well as the AI caption available. This approach ensured a balanced and comparable set of human and AI texts per image, resulting in 101 images × 12 human captions = 1,212 human captions, plus 101 AI captions, for a total of 1,313 texts. This subset ("main") was used for all exploratory analyses of linguistic features and for training classification models.

To create a held-out evaluation set, we randomly selected 100 images from the remaining dataset. For each image, we extracted both the AI-generated description and one randomly chosen human caption, resulting in 200 texts evenly balanced between the two sources (AI and human). This dataset, referred to as "paired", was used to compare results obtained from the main dataset and to test classifiers on unseen material (see Table I).

TABLE I. DATASET COMPOSITION AND DESCRIPTIVE STATISTICS

Dataset	Size (texts)	AI texts	Human texts	Feature	AI mean (SD)	Human mean (SD)
Main	1,313	101	1,212	n_words	64.3 (21.7)	54.0 (31.4)
				n_chars	449.6 (155.6)	354.0 (207.8)
Paired	200	100	100	n_words	66.6 (22.9)	64.2 (44.3)
				n_chars	472.0 (174.8)	424.3 (287.6)

In the **main dataset**, AI-generated captions were consistently longer than human ones, with significant differences in both word and character counts (small–moderate effect sizes, $p < .001$). In the **paired dataset**, by contrast, no reliable length differences emerged between AI and human captions.

Variance analysis showed that in the paired dataset, human captions exhibited significantly greater variability than AI captions, both in word count (Levene's test, $F(1,198) = 21.23$, $p < 0.00001$) and character count ($F(1,198) = 15.42$, $p < 0.001$). In contrast, the main dataset revealed no significant differences in variance for character count ($F(1,1311) = 2.46$, $p = 0.117$). However, human texts in the main dataset exhibited greater variance in word count compared to AI (Levene's test, $F(1,198) = 5.57$, $p < 0.018$). This indicates that individual human captions can vary widely in length (at least in terms of word count), while AI-generated captions tend to be more consistent.

To prevent subsequent analyses from being simply influenced by differences in text length, we control for potential effect of text length (measured in word count) when selecting linguistic features.

B. Features

We developed a feature set grounded in previous studies examining systemic differences between AI-genereated and human texts [6-10], as well as our experience with various authorship analysis tasks [12]. The features encompass (1) lexical diversity, (2) morphosyntactic density and complexity, (3) syntactic roles, and (4) semantic concept engagement.

1) Lexical diversity

Since our captions are relatively short, we prioritize indices known to be more stable for short texts [13-15] and statistically control for any residual length effects.

We calculated three well-established indices that are robust for short texts:

1. **HDD (Hypergeometric Distribution Diversity)** was calculated manually following the method described by McCarthy & Jarvis [14, 15]. HDD estimates, for each word type, the probability that it appears at least once in a random sample of tokens of size *s* from the text, averaged across all word types. We compute HD-D using the standard sample size s=42 [15]. For captions shorter than 42 tokens, we set s=min(42, N−1) to comply with the hypergeometric formulation and avoid the degenerate case where s=N. We use s=42 as the de facto standard in the HD-D literature to maintain comparability across studies and languages. A higher HDD value indicates greater lexical diversity.

2. **MATTR (Moving-Average Type-Token Ratio)** was calculated using the implementation in **quanteda.textstats** R package, with the default window size set to 50 tokens [16]. MATTR computes the type-token ratio within a fixed-length moving window and averages the results across the entire text. For texts shorter than 50 tokens, we reduced the window size to a minimum of 10 tokens to ensure full coverage of all documents. This approach helps stabilize the type-token ratio against text length effects, where higher values indicate greater lexical diversity [13].

3. **Maas** (**Maas index)** was calculated using **quanteda.textstats**. This measure applies a logarithmic transformation of the type-token ratio, which reduces sensitivity to text length. Unlike MATTR and HDD, lower Maas values indicate greater lexical diversity, whereas higher values reflect increased lexical repetition. Therefore, the Maas index complements MATTR and HDD by offering a length-adjusted estimate of lexical variety on a different scale [13].

All three indices were calculated using tokenized texts that were lowercased, with punctuation and numbers removed, and stopwords filtered out to highlight the content vocabulary.

We conducted two sets of calculations: 1) a raw pipeline where tokens were surface forms; and 2) a lemmatized pipeline where tokens were lemmas extracted using the udpipe R package with the pretrained Russian model.

This dual approach enabled us to assess whether the observed differences in lexical diversity were influenced by inflectional variation (surface forms) or by normalized lexical forms (lemmas).

To account for the well-established dependence of lexical diversity indices on text length, we combined two length-control strategies: 1) residualization, where each metric was regressed on the logarithm of token count, followed by Welch's t-tests on the residuals to compare AI and human captions; and 2) ANCOVA, using raw indices served as outcomes in models that include both caption source (AI vs. human) and log token count as predictors. Residualization offers a model-free approach to length control for simple group comparisons, while ANCOVA provides adjusted effects and interpretable adjusted means. Convergence between the residualized and ANCOVA results increases confidence in the robustness of our findings. To control for multiple testing, the Benjamini–Hochberg FDR correction was applied within each analysis type (residualized vs. ANCOVA) across the three indices per dataset.

The same length-control strategies were consistently applied to other feature sets beyond lexical diversity as well.

2) Morphosyntactic density and complexity

We calculated a set of morphological and syntactic features to capture the structural properties of captions authored by AI and humans. These measures targeted both the composition of morphological categories and the complexity of syntax [17-19]. Specifically, they included:

Morphological composition:

– **upos_entropy**: normalized Shannon entropy of Universal POS (UPOS) tag distributions (**POS_entropy**). Higher values indicate a more balanced mix of parts of speech;

– **upos_bigram_entropy**: normalized Shannon entropy of adjacent UPOS tag bigrams, capturing the diversity of sequential POS (**POS_bigram_H**). Higher values indicate greater variation in POS sequencing, suggesting less formulaic language;

– **func_ratio**: the proportion of function words relative to the total number of tokens. Higher values indicate a greater presence of functional scaffolding compared to content words.

Morphological variety:

– **morph_entropy (H)_<feature>:** normalized entropy of values within specific morphological categories (e.g., H_Case_N, H_Gend_N, H_Anim_N, H_Tense_V, etc.). Following [18], the entropy of morphological categories provides a reliable measure of how evenly grammatical distinctions are distributed throughout a text. Higher entropy values indicate a more balanced use of the category's values, reflecting greater morphological diversity (e.g., a balanced use of several noun cases or a mix of animate and inanimate nouns). In contrast, lower values suggest that one or a few values dominate, indicating limited variety (such as nearly all nouns appearing in the nominative case).

– **feat_per_token:** mean number of marked morphological attributes per token indicates the richness of morphological marking. Higher values indicate a more complex morphology, where tokens typically encode multiple grammatical distinctions simultaneously. This usually corresponds to the presence of inflected nouns, verbs, and adjectives carrying several features at once. In contrast, lower values suggest lighter morphological marking, with texts containing more uninflected words (such as function words, particles, and short adverbs) or primarily analytic structures.

Inflectional richness:

– **SD3_NOUN / SD3_VERB / SD3_ADJ:** a set of features that capture the diversity of suffixes within three major inflecting word classes (nouns, verbs, and adjectives). For each token, the last three characters of its surface form are extracted to approximate morphological endings in Russian, where inflection is primarily expressed through suffixes. Formally, it is defined as the number of distinct three-character endings observed in a document, normalized by the total number of tokens in that word class. Higher values indicate greater morphological productivity or variability, meaning the text uses a wider range of inflected forms (e.g., nouns appearing in multiple cases or verbs across different tenses, aspects, or persons). Lower values indicate restricted or repetitive suffix patterns, where fewer endings dominate, suggesting limited inflectional variation (e.g., mostly nominative singular nouns or predominantly infinitive verbs)**.**

– **AFL (Average Forms per Lemma):** mean number of distinct surface forms per lemma. Higher values suggest a richer inflectional variety for each lemma, meaning that lemmas occur in multiple paradigmatic forms. In contrast, lower values indicate more limited inflection, thus quantifying morphological complexity by the extent of form proliferation.

– **BFL (Average Bundles per Lemma):** mean number of unique morphological bundles (combinations of features) realized per lemma. Each bundle is defined by a set of morphological attributes in the UD annotation (e.g., *Case=Gen, Number=Plur, Gender=Fem*). While multiple surface forms may correspond to the same bundle (e.g., spelling variants), BFL counts only the unique bundles. Higher values indicate that lemmas appear in a wider range of morphosyntactic contexts. This index captures functional inflectional diversity by abstracting from spelling variations and focusing on grammatical categories.

– **Inflectional Diversity Index (IDI)** is a ratio-based measure that quantifies the number of morphological bundles realized relative to the lemma inventory. Essentially, it is effectively a normalized version of BFL. Higher IDI values indicate that lemmas are expressed through a greater variety of distinct inflectional bundles, reflecting a broader functional range. While BFL calculates the average number of bundles per lemma, IDI presents this information as a ratio, making it easier to compare across corpora or texts with different lemma sizes.

– **Morphological Bundle Type–Token Ratio (MB_TTR)** is a type–token ratio calculated not only on word forms but on lemma–bundle pairs (e.g., стол + Case=Gen.Sg; стол + Case=Nom.Pl). Similar to traditional TTR, higher MB_TTR values indicate greater variety, but with the added morphological information.

– **editdist_mean (Mean Lemma–Token Edit Distance)** is the average Levenshtein distance between each surface token and its corresponding lemma, measuring the morphological difference from the base form. Higher values indicate greater inflectional or derivational complexity (e.g., long case or tense suffixes, stem alternations).

Lemma distributional shape:

– **hapax_share** is the proportion of lemmas that appear only once in a document (known as hapax legomena);

– **gini_lemma** is the Gini coefficient computed over the lemma frequency distribution, reflecting the inequality of lemma reuse. Higher values indicate dominance by a small set of lemmas, while low value reflect a more equal lemma distribution.

Both metrics are computed using lemmas (after lemmatization) and thus correspond directly to our morphological features. Unlike HDD/MATTR, which assess global lexical diversity through type-token dynamics, these measures target the distributional shape of lemmas, specifying long-tail variety and inequality in reuse.

Syntactic complexity:

– **depdist_mean (Mean Dependency Distance)** is the average linear distance between a head and its dependent within a syntactic dependency tree, excluding root relations (Liu 2008). Higher values indicate longer distances and, consequently, greater syntactic complexity**;**

– **verb_sent_mean (Verbal Density per Sentence)** represents the average number of verbs per sentence. Higher values indicate that sentences contain more predicates or verbal events, reflecting a denser clausal structure and potentially more complex event packaging.

– **subord_ratio (Subordination Ratio)** is the proportion of sentences containing at least one subordinate or clausal dependent relation, such as adverbial clause, clausal complement, open clausal complement, adnominal clause, or relative clause. A higher presence of these structures increases syntactic dependency and complexity, while lower values indicate more paratactic or simpler sentence structures.

All morphosyntactic features were computed on texts annotated using the UDPipe Russian model. Punctuation and symbols were excluded. To avoid undefined values, we set log(tokens)=log(1) when token counts were zero. For metrics requiring a specific part-of-speech category (e.g., entropy of verb tense, noun case, or subject/object counts), documents lacking the relevant POS was absent were assigned *NA* for that metric and excluded pairwise from statistical testing. Normalization and statistical controls for length followed our pipeline: residualization of each metric against the log token count and ANCOVA with log length as a covariate.

3) Syntactic Roles

We quantified core syntactic roles using Universal Dependencies (UD) relations from the UDPipe annotation of the Russian captions. For each token, we extracted its dependency relation (dep_rel) and counted occurrences of the following UD labels, including all subtypes: nsubj (nominal subject), obj (direct object), iobj (indirect object), amod (adjectival modifier), obl (oblique nominals, e.g., locatives, instrumentals), and conj (conjuncts, i.e., coordinated elements linked to their head conjunct). Additionally, we created a combined object count: obj_any = obj + iobj, to capture the total object realization regardless of direct or indirect status.

To make counts comparable across documents of varying lengths, we converted raw counts to rates per 100 tokens.

4) Semantic features

We present a novel set of semantic features based on concept engagement functionality, as implemented in the text2map R package [20]. This approach estimates the extent to which a text engages with specific semantic concepts by calculating the optimal transport cost required to map a document's lexical content onto predefined concept prototypes (lower cost indicates stronger alignment, see also [21, 22] for details).

Building on prior research into the componential structure of word meaning in language models and recent neuroscientific studies on the representation of semantic components in the brain [23], we compiled a list of 46 semantic features describing word meanings [24]. These features are grouped by major modalities and domains, including perceptual (visual, auditory, somatic, gustatory, olfactory), motor and spatiotemporal dimensions, affective-emotional, cognitive and social domains, and drives.

Each concept prototype was defined as the centroid of its exemplar word vectors, and the alignment of a text with the prototype was computed using vector similarity. Oppositional dimensions were constructed by defining semantic axes (e.g., concrete–abstract, pleasant–unpleasant, static–dynamic). Texts were scored based on their projection onto these axes.

Concept prototypes and semantic scales were derived from multiple sources. For example, to develop features reflecting the salience of the visual modality in text, we created a semantic scale with one pole consisting of words rated highest on and the opposite pole consisting of words rated lowest as reported in [25]. Additionally, we utilized resources we previously developed [26] and exemplar lists provided by [23]. We also included a core set of semantic-differential oppositions [27]. High values on these features indicate stronger alignment of the text with the corresponding semantic domain. For instance, a high **VisNorms** score reflects frequent text engagement with the visual modality; a high **EmoHappy** score indicates a stronger presence of positive affect; and a high **CognitionAbstract** score corresponds to more abstract, conceptual vocabulary.

The resulting 46-dimensional semantic feature set has been successfully applied in our prior research on authorship profiling [12] and in modelling individual differences in the word meaning [24, 27]. These features provide insight into which conceptual systems (sensory, cognitive, emotional, etc.) are more salient in AI versus human captions.

C. Classification

We evaluated the predictive value of the linguistic features by training binary classifiers to distinguish between AI-generated and human-generated captions. Models were trained on the main dataset and tested on a held-out paired set of 100 unseen images, each containing one AI captions and one human caption. Model selection employed grouped cross-validation with groups defined by image_id, ensuring that captions from the same image did not appear in both training and validation folds. We used three classifiers: elastic net logistic regression (our primary baseline, combining L1 and L2 regularization), ridge logistic regression (L2 regularization only; linear, interpretable comparator), and a Support Vector Machine (SVM) with a radial basis function (RBF) kernel (a non-linear comparator).

Elastic Net Logistic Regression was chosen as the primary model because it performs embedded feature selection, handles correlated predictors effectively, and yields well-calibrated probabilities for interpretability. Ridge Logistic Regression was included as a stability check, since it applies shrinkage without variable selection, thereby testing whether the same decision boundary is recovered without sparsity.

From the elastic net model at its selected penalty, we collected features with non-zero coefficients, and then removed redundancy using a de-correlation filter (Pearson $\rho > 0.98$), retaining the strongest coefficient in each cluster. The retained subset was re-fitted using bias-reduced (Firth) logistic regression to obtain odds ratios (ORs) and 95% CIs. ORs are reported for the AI label (OR > 1 increases odds of AI; OR < 1 increases odds of Human).

IV. Results

Table II reports adjusted effect estimates (est) and FDR-corrected *p*-values for raw-form and lemma-based lexical diversity indices, across both the main and paired datasets.

TABLE II. Results of ANCOVA models for lexical diversity indices (HDD, MATTR, MAAS) in AI vs. human captions, adjusted for log token count

Dataset	Metric	Raw est	Raw FDR	Lemma est	Lemma FDR
Main	HDD	-2.621	0.000	-2.767	0.000
Main	MATTR	-0.001	0.154	-0.003	0.038
Main	Maas	0.012	0.038	0.016	0.009
Paired	HDD	-4.530	0.000	-4.320	0.000
Paired	MATTR	-0.002	0.061	-0.004	0.001
Paired	Maas	0.010	0.124	0.006	0.399

HDD values were significantly higher for AI across both raw and lemmatized pipelines in both datasets, indicating a robust breadth of vocabulary. MATTR also showed a similar trend, although statistical significance was observed only for the lemmatized texts in both datasets. This suggests that AI outputs maintain a broad and evenly distributed vocabulary relative to length, particularly when inflectional variation is controlled.

By contrast, the Maas index which penalizes repetition more heavily and is sensitive to low-frequency types identified greater repetitiveness in human texts, but only in the larger main dataset. Even after lemmatization, AI-generated captions employed a wider range of rarer lemmas and exhibited less formulaic repetition, resulting in lower Maas values. The absence of this effect in the smaller paired dataset likely reflects greater human variation on this metric.

Overall, converging evidence from HDD, MATTR, and Maas consistently indicates that **AI-generated descriptions exhibit greater lexical diversity than human-generated descriptions.**

Syntactic complexity features were the strongest discriminators between AI- and human-authored captions. Across both datasets, the subordination ratio, verbs per sentence, and mean dependency distance exhibited large, highly significant effects (all FDR-corrected $p < 10^{-6}$) (Fig. 1-2). AI captions contained more verbs per sentence (main: +0.47; paired: +0.66 relative to human), a markedly higher proportion of subordinate clauses (main: +0.24; paired: +0.36), and longer average dependency distances (main: +0.24; paired: +0.31). These patterns indicate that **AI outputs are structurally more complex, constructing multi-clausal sentences with deeper embedding and longer head–dependent spans**.

By contrast, human captions demonstrated a modest but consistent advantage in inflectional variety per lemma. Average Forms per Lemma (AFL), Average Bundles per Lemma (BFL) and Inflectional Diversity Index (IDI) were slightly higher in human texts (approximately +0.02–0.03), indicating that although **human captions are shorter and syntactically simpler, they exhibit somewhat richer paradigmatic variation across lemmas**.

Analysis of morphological entropy indices further refines this picture: AI captions exhibited greater diversity in case and gender marking, consistent with their more elaborate noun phrase constructions. Human captions varied the number, person, and animacy slightly more.

AI-generated captions tend to exhibit more morphological features per token, reflecting their higher structural complexity. This highlights a key difference: AI captions build complexity through clausal embedding and morphologically dense tokens, while human captions achieve variety by employing a wider range of lemma-level inflectional forms. Rather than a contradiction, this difference reflects a trade-off between local morphological density and paradigmatic breadth. AI-generated texts often feature longer, more clause-heavy sentences, with individual words carrying multiple morphological features (e.g., verbs marked for tense, aspect, person, number or adjectives with case–gender–number agreement). This increases the feature load per token. In contrast, human captions tend to be syntactically simpler but exhibit a slightly broader paradigmatic range, reusing the same word roots in a more diverse set of forms and feature bundles. In summary, AI achieves complexity through structural embedding and locally

dense morphology, whereas humans achieve complexity through lemma-level inflectional variety.

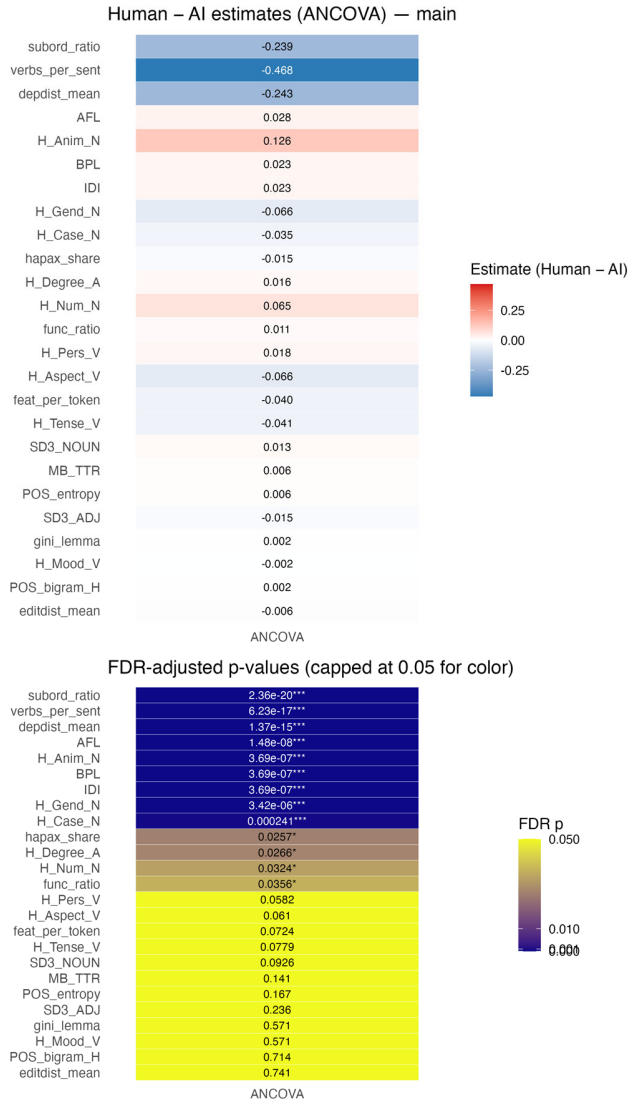


Fig. 1. Differences in morphosyntactic features between human and AI captions in the main dataset (ANCOVA estimates adjusted for log token count). Top panel: Human–AI estimate contrasts (positive = higher in Human; negative = higher in AI). Bottom panel: Benjamini–Hochberg FDR-adjusted p-values

Syntactic role analysis further highlights the structural complexity of AI-generated captions (Fig. 3). In the main dataset, AI texts contained more objects per 100 tokens - both direct objects (obj, −0.76, pFDR = .0035), indirect objects (iobj, −0.57, pFDR < .001), and their combined total (obj + iobj, −1.32, pFDR < .001). AI also produced significantly more coordination structures (conj, −1.01, pFDR = .0015). In contrast, human captions exhibited richer nominal packaging, with more adjectival modifiers (amod, +0.99, pFDR = .058, trend-level) and more obliques (obl, +1.07, pFDR = .007). The rate of subjects (nsubj) did not differ between groups (−0.11, pFDR = .72).

The paired dataset replicated the ***object pattern***: AI again used more direct objects (−0.85, pFDR = .016), indirect objects (−0.80, pFDR < .001), and their combined total (−1.65, pFDR < .001). Human captions once again favored adjectival modification (amod, +1.86, pFDR = .012). Differences in coordination (conj, −0.85, pFDR = .084) and obliques (obl, +0.30, pFDR = .58) were not statistically significant in this smaller sample. Subjects showed no consistent differences (nsubj, −0.51, pFDR = .24).

Taken together, these results confirm that AI-generated captions achieve complexity through object density and coordination, whereas human captions emphasize nominal elaboration by using modifiers and obliques.

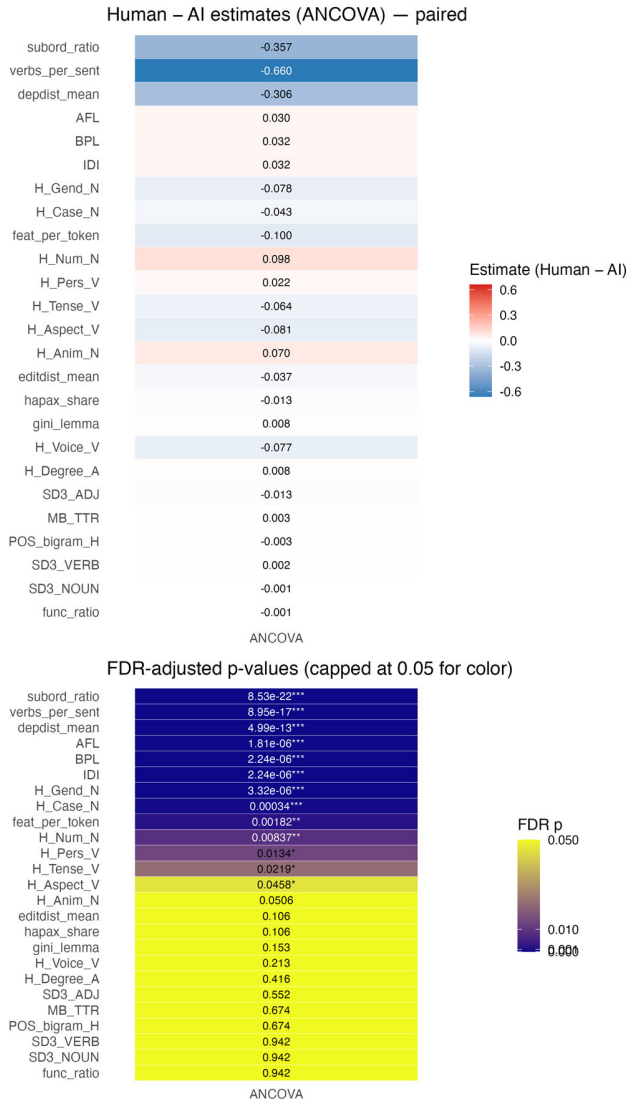
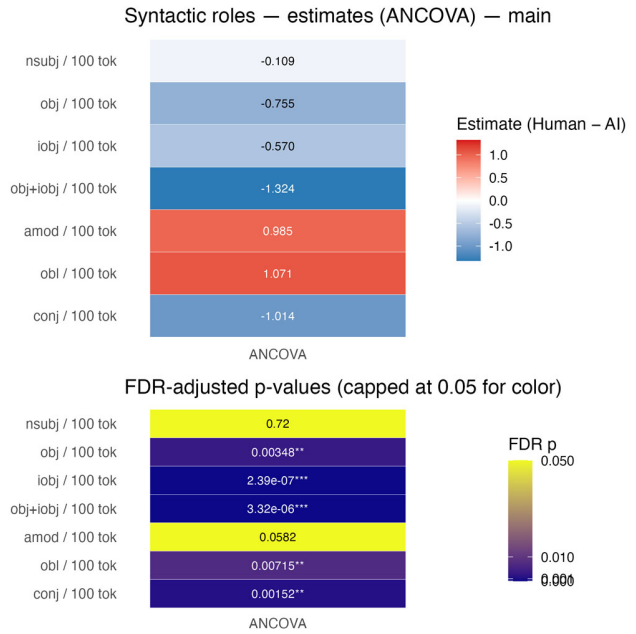


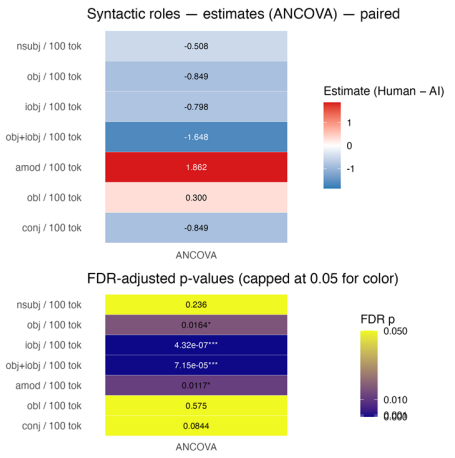
Fig. 2. Differences in morphosyntactic features between human and AI captions in the paired dataset (ANCOVA estimates adjusted for log token count). Top panel: Human–AI estimate contrasts (positive = higher in Human; negative = higher in AI). Bottom panel: Benjamini–Hochberg FDR-adjusted p-values

Overall, the object effect is consistently replicated across both datasets as the most stable AI grammar signal. AI style is predicate-heavy, featuring more objects and often enumerative, with increased coordination. Humans prefer noun-centric descriptions, conveying detail through adjectives and

prepositional phrases (obliques) that add scene setting and attributes.



a)



b)

Fig. 3. Differences in syntactic roles preference between human and AI captions in the main (a) and paired (b) dataset (ANCOVA estimates adjusted for log token count). Top panel: Human–AI estimate contrasts (positive = higher in Human; negative = higher in AI). Bottom panel: Benjamini–Hochberg FDR-adjusted p-values

Semantic feature analysis revealed a robust and consistently replicated distinction between AI- and human-authored captions across both the main and paired datasets (Fig. 4). This pattern was observed with large effect sizes in categories reflecting the opposition between abstract and concrete language (**CognitionAbstract**, main dz = 1.88; paired dz = 1.66; **CognitionLIWC**, 1.34; 1.24), indicating strong AI engagement with abstract words and terms related to cognitive processes.

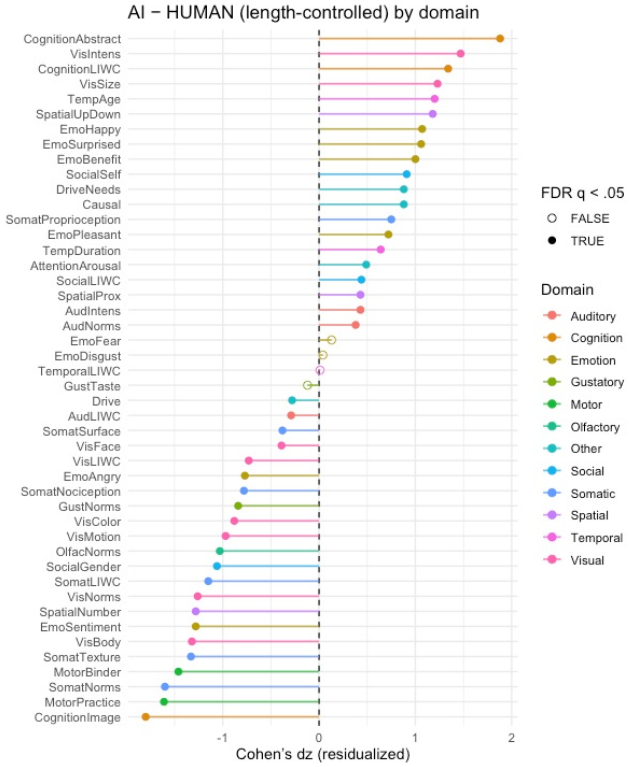


Fig. 4. Differences between AI- and human-authored captions across semantic domains. Positive values indicate higher scores in AI captions; negative values indicate higher scores in human captions.

AI texts are characterized by higher values of opposition poles in brightness (bright vs. dark: **VisIntens** 1.47 vs. 1.37), size (large vs. small: **VisSize** 1.23 vs. 1.13), age (old vs. new: **TempAge** 1.20 vs. 1.05), and spatial orientation (vertical vs. horizontal framing: **SpatialUpDown** 1.18 vs. 1.14).

Other domains also consistently skewed toward AI, including positive and surprising emotions (**EmoHappy**, **EmoSurprised**), usefulness (**EmoBenefit**, **DriveNeeds**), causality (**Causal**), and words describing socially approvable personality traits such as *kind* and *talented* (**SocialSelf**). Higher values in contrasts like long–short (**TempDuration**), loud–quiet, and close–far also characterize AI captions. Overall, AI-generated texts are more engaged with the auditory modality (**AudIntens**, **AudNorms**). These effects reinforce the impression of AI captions as more interpretive, abstract, and evaluative, frequently employing degree, scale, and causal framings, as well as positive affect.

In contrast, human captions relied more on concrete, embodied, and perceptually grounded language. Strong replicated effects were observed for concrete words (**CognitionImage**: main dz = −1.80; paired = −1.62), words related to motor modality and objects with which one could have personal experience (**MotorPractice**: −1.61; −1.46; **MotorBinder**: −1.46; −1.18; **VisMotion**: −0.97; −0.86), somatosensory descriptors (**SomatNorms**: −1.60; −1.43; **SomatTexture**: −1.33; −1.20; **SomatLIWC/Nociception/Surface**: −0.6 to −1.15), descriptions of body and face terms (**VisBody**: −1.32; −1.10; **VisFace**: −0.3 to −0.73), visual modality overall (**VisNorms**:

−1.26; −1.08), colors (**VisColor**: −0.88; −0.75), numbers (**SpatialNumber**: −1.28; −1.05), olfactory and gustatory terms (**OlfacNorms**: −1.03; −1.08; **GustNorms**: −0.84; −0.83; **GustTaste**: paired only, −0.21, q = .038), and emotionally charged words (**EmoSentiment**: −1.28; −1.13).

Thus, human captions were more grounded in perceptual experience, frequently referencing body parts, colors, movements, textures, tastes, smells, emotions, as well as richer motor and somatosensory language.

Crucially, these effects cannot be explained by caption length. Correlations between AI–Human semantic score differences and token-count disparities were modest (median $|r| \approx 0.18$), accounting for no more than 7% of the variance. After regressing out caption length, effect sizes attenuated only minimally (median $|\Delta dz| \approx$ a few hundredths). This demonstrates that the semantic divergence between AI and human captions reflects genuine differences in content selection and framing, rather than verbosity.

In summary, AI-generated captions tend to emphasize abstract, scalar, and affective aspects of descriptions, while human-generated captions focus more on embodied, perceptual, and concrete experiences. This distinction reflects a fundamental stylistic contrast: AI leans toward interpretive commentary, whereas humans remain closer to the sensory details of the depicted scene.

We evaluated three classifiers representing complementary modeling strategies (Table III).

TABLE III. CLASSIFICATION PERFORMANCE ON THE TEST DATASET

Classifier	BAcc	Precision	Recall	F1
Elastic Net Logistic Regression	0.9	0.955	0.84	0.894
Ridge Logistic Regression	0.9	0.955	0.84	0.894
Support Vector Machine (RBF)	0.875	0.844	0.92	0.88

Both linear models achieved balanced accuracy of 0.900, with high precision (0.955) and good recall (0.840), confirming their consistency. The confusion matrix (84/100 AI and 96/100 human captions correctly classified) illustrates this balance, showing that linear models preserved strong recall while maximizing precision.

To contrast with linear approaches, we also tested a nonlinear Support Vector Machine (RBF kernel). This classifier reached higher recall for AI (0.920) but at the expense of lower precision (0.844), yielding slightly reduced balanced accuracy (0.875).

Together, these results indicate that linear penalized logistic regression provides both high accuracy and stable calibration, while the SVM demonstrates that nonlinear decision boundaries can shift the precision–recall trade-off but do not improve overall generalization.

While classification results demonstrate that AI and human captions can be reliably distinguished, they do not reveal which linguistic features drive the distinction. To address this, we applied bias-reduced logistic regression to the de-correlated elastic-net subset. This second-stage analysis provided interpretability beyond performance metrics and identified seven predictors with odds ratios significantly different from 1.0, confirming their robust contribution to AI–human discrimination. Features associated with AI captions (OR > 1) included **subordination ratio** (OR 67.0), **mean dependency distance** (OR 11.5), **CognitionLIWC** (OR 2.66), **lexical diversity** (HDD_Lemma) (OR 1.06). Features associated with human captions (OR < 1) were **Attentional arousal** (OR 0.46), **SomaticLIWC** (OR 0.38), **SomatNociception** (OR 0.24).

AttentionArousal is higher in AI on average, but after accounting for stronger, correlated predictors (e.g., subordination, dependency distance, cognition), its unique contribution predicts Human (conditional OR < 1), a standard suppression effect due to feature overlap.

V. DISCUSSION

Our results converge across lexical, morphosyntactic, and semantic dimensions, revealing a systematic stylistic difference between AI- and human-authored captions.

AI captions consistently exhibited greater diversity according to HDD and MATTR metrics, while Maas metric also indicated lower repetition in the main dataset. These findings suggest that AI-generated texts utilize a broader and more evenly distributed vocabulary compared to human-authored texts.

At the structural level, AI-generated captions exhibited greater syntactic complexity, characterized by more verbs per sentence, an increased number of subordinate clauses, and longer dependency distances. They also included more direct and indirect objects, along with greater coordination, indicating denser propositional content. In contrast, human captions favored nominal elaboration, with more adjectival modifiers and oblique dependents, and demonstrated a small but consistent advantage in lemma-level inflectional variety. This reflects a trade-off: AI complexity arises through clausal embedding and morphological density per token, whereas human complexity emerges from the paradigmatic diversity of inflection across lemmas.

The semantic feature analysis revealed a complementary division of labor. AI-generated captions were enriched with abstract, scalar, and interpretive language featuring abstract cognition, intensity and size descriptors, causal framing and expressions of positive or surprising emotions. In contrast, human-generated captions emphasized concrete, perceptual, and embodied content, including visual details (such as color, motion, faces, and bodies), somatosensory descriptors, motor actions, and olfactory and gustatory terms. Thus, while AI tends to describe images in an interpretive and evaluative manner, humans remain more closely connected to situated sensory experience.

This contrast carries both methodological and theoretical implications. Methodologically, it demonstrates that semantic feature sets can effectively capture the nuanced differences between AI and human language, offering interpretable evidence of distinctions that remain undetected when relying solely on lexical or syntactic measures. Our results suggest that AI-generated descriptions tend to replicate the higher-level

interpretive aspects of meaning-making, while human descriptions highlight the sensory foundations of experience.

Crucially, our predictor analysis mirrors some published linguistic patterns: AI captions show greater structural complexity (higher subordination ratio; longer dependency distances), more cognition-oriented vocabulary, and higher lexical diversity, whereas human captions exhibit more embodied/attentional content. This convergence suggests that the discriminative cues reported in earlier work generalize beyond expository essays to visual description in a different language and register.

Importantly, we controlled for text length, a confounding factor often overlooked in prior studies, and found that the AI-Human contrasts remained robust even after adjusting for text length. This suggests that the observed differences are not merely artifacts of verbosity but reflect genuine stylistic distinctions. Despite the shift to multimodal image description, our classifiers and feature-level signals closely track patterns reported in the literature, and the length-controlled evaluation indicates these effects reflect stylistic/structural properties of AI vs. human texts – not quirks of the captioning task.

By placing our findings within the broader literature, we demonstrate that linguistic expertise is essential for interpreting the outputs of LLMs. While previous research on LLMs has often prioritized task performance over linguistic analysis, our results highlight the value of examining lexical, morphosyntactic and semantic features. Additionally, whereas most prior studies focus on English, our work offers a new perspective by analyzing Russian image descriptions, thereby broadening the empirical foundation for stylistic evaluation.

To our knowledge, this is the first study to examine linguistic characteristics of MLLM-generated captions in direct comparison with human descriptions. This research contributes to the field by demonstrating that the stylistic divergences observed in text-only LLMs also apply to multimodal contexts, where challenges related to perceptual grounding and interpretive abstraction are especially prominent.

Several limitations of this study should be noted. First, our analysis was limited to Russian-language captions, so cross-linguistic generalizability remains to be established, particularly given typological differences in morphology and syntax. Second, we evaluated AI outputs from only one MLMM, and results may differ across various model architectures or training approaches. Third, we inspected only one type of image (artworks). We therefore treat our paired Russian caption dataset as a controlled testbed with strong internal validity, while leaving cross-domain external validity as an open question. Future work should stress-test robustness across languages, registers (captions vs. essays/news), and model families (vision-language vs. text-only LLMs), and probe whether the same structural complexity and cognition-lexicon signals persist when image types or prompt styles vary.

VI. Conclusion

We developed a comprehensive pipeline to compare Russian AI-generated captions with human-generated ones across linguistic, syntactic, content, and semantic dimensions, uncovering systematic differences between them.

The implications of this work span a wide range of applied fields. Beyond simply detecting objects or suggesting diagnoses, the safety and effectiveness of vision-language systems depend on how they generate text – whether descriptions are concrete or abstract, perceptual or cognitive, and evaluative or descriptive; whether they accurately handle negation, uncertainty, and spatial language; whether their syntax remains clear and well-grounded. A feature-centric analysis anchored in our semantic oppositions and complemented by lexical, morphological and syntactic features makes model behavior auditable, could uncover failure modes (such as overconfidence and bias) and enables targeted control. Given the growing importance of prompt design, especially in sensitive domains such as medicine, future work should systematically examine how different prompt formulations shape the balance between abstract/interpretive and concrete/perceptual language. This issue is particularly relevant for emerging imaging modalities such as laser-induced contrast visualization (LICV) [28-29], an optical technique that uses laser excitation to generate high-contrast, label-free images of biological tissues. LICV offers complementary structural and functional information compared to conventional imaging methods (e.g., X-ray, CT), and its adoption will require AI captioning systems that are both accurate and clinically interpretable. In clinical or diagnostic settings, overly interpretive or evaluative AI-generated captions could mislead practitioners, while insufficiently descriptive outputs might omit critical details.

Beyond their value for distinguishing AI- and human-authored texts, the linguistic contrasts identified in this study – syntactic complexity, lexical diversity, and semantic oppositions between cognition-related and embodied/attentional vocabulary – are also highly relevant to second language acquisition. These features capture the kinds of structural and semantic contrasts that often pose challenges for learners of Russian as a foreign language. By showing that such dimensions can be reliably extracted and quantified, our approach provides a methodological foundation for developing pedagogical tools and resources that emphasize functionally meaningful contrasts in Russian usage.

Acknowledgment

Tatiana Litvinova and Viktoriya Molchanova acknowledge the support of the Ministry of Education of the Russian Federation within the framework of the state task in the field of science (topic number QRPK-2025-0013). Pavel Seredin and Olga Dekhnich did not receive funding for their work on this paper.

We are grateful to the dataset creators, Anastasiya Kolmogorova and Polina Nalobina, for providing the access to the dataset.

References

[1] J. Zhang, D. Zhang, X. He, F. Gao and X. Li. "An Overview of Image Captioning Generation Based on Large Language Models", in: El Yacoubi, M.A. (eds). *Trends in Haptics and Virtual Reality. ICHVR*

2024. Learning and Analytics in Intelligent Systems, vol. 50, 2025, Springer, pp 50–56.

[2] S. Sarto, M.Cornia and R. Cucchiara, "Image Captioning Evaluation in the Age of Multimodal LLMs: Challenges and Future Perspectives", *arXiv,* 2025, 2503.14604.

[3] K. Cheng, W. Song, J. Fan, Z. Ma, Q. Sun, F. Xu, C. Yan, N. Chen, J. Zhang and J. Chen, "CapArena: Benchmarking and Analyzing Detailed Image Captioning in the LLM Era", in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, p. 14077–14094.

[4] J. Ji, Y. Hou, X. Chen, Y. Pan and Y. Xiang, "Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study", *JMIR Form Res*, 2024, Feb 8;8:e32690.

[5] R. Agbareia, M. Omar, O. Zloto, B.S. Glicksberg, G.N. Nadkarni and E. Klang, "Multimodal LLMs for retinal disease diagnosis via OCT: few-shot versus single-shot learning", *Therapeutic Advances in Ophthalmology*, 2025, 17.

[6] A. Reinhart, D.W.Brown, B. Markey, M. Laudenbach, K. Pantusen, R. Yurko and G. Weinberg, "Do LLMs write like humans? Variation in grammatical and rhetorical styles", *Proceedings of the National Academy of Sciences of the United States of America,* 2024, vol. *122(8)*.

[7] B. Markey, D. W. Brown, M. Laudenbach and A. Kohler, "Dense and disconnected: Analyzing the sedimented style of ChatGPT-generated text at scale", *Writ. Commun.* 2024, 41, pp. 571–600.

[8] F. Jiang and K. Hyland, "Does ChatGPT Argue Like Students? Bundles in Argumentative Essays", *Applied Linguistics*, 2024, vol. 46(3), pp. 375-391.

[9] A. Muñoz-Ortiz, C. Gómez-Rodríguez and D. Vilares, "Contrasting Linguistic Patterns in Human and LLM-Generated News Text", *Artif Intell Rev*, 2024, vol. 57, 265.

[10] M. Sandler, H. Choung, A. Ross and P. David, "A Linguistic Comparison between Human and ChatGPT-Generated Conversations", *ArXiv*, 2024, abs/2401.16587.

[11] A. Kolmogorova and P. Nalobina, "Database of textual descriptions and vector embeddings of artworks from Digital Collection of the Hermitage", certificate of registration, 2025, No. 6.0018-2025.

[12] T. Litvinova, V. Zavarzina and P. Panicheva, "Individual Differences in the Most Frequent Content Word Usage as a New Type of Features in the Authorship Profiling Task", in *Proceedings of The International Society for Technology Education and Science, ICRES 2024*, 2024, p. 1901–1924.

[13] F. Zenker and K. Kyle, "Investigating minimum text lengths for lexical diversity indices", *Assessing Writing*, 2021, vol. 47, 100505.

[14] P. M. McCarthy and S. Jarvis, "vocd: A theoretical and empirical evaluation", *Language Testing*, 2007, vol. 24, pp. 459–488.

[15] P.M. McCarthy and S. Jarvis, "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment", *Behavior Research Methods*, 2010, vol. 42, pp. 381–392.

[16] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller and A. Matsuo, "quanteda: An R package for the quantitative analysis of textual data", *Journal of Open Source Software*, 2018, vol. **3**(30), 774.

[17] M. Bane, "Quantifying and measuring morphological complexity", in *Proceedings of the 26th WCCFL,* 2008.

[18] B. Bickel, "Distributional Typology: Statistical Inquiries into the Dynamics of Linguistic Diversity", in Bernd Heine, and Heiko Narrog (eds), *The Oxford Handbook of Linguistic Analysis*, 2nd edn, 2015.

[19] H. Liu, "Dependency distance as a metric of language comprehension difficulty", *Journal of Cognitive Science,* 2008, vol. *9*(2), pp. 159–191

[20] D.S. Stoltz and M.A. Taylor, "text2map: R Tools for Text Matrices", *Journal of Open Source Software*, 2022, vol. 7(72):3741.

[21] M.A.Taylor and D.S. Stoltz, "Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement", *Journal of Computational Social Science*, 2021, vol. 4, pp. 231–242.

[22] D. S. Stoltz, M. A.Taylor, J. S. K. Dudley, "A Tool Kit for Relation Induction in Text Analysis", *Sociological Methods & Research*, 2024, vol. 4 (2), pp. 565-604.

[23] J.R. Binder, L.L. Conant, C.J. Humphries, L.F ernandino, S. B. Simons, M. Aguilar, "Toward a Brain-Based Componential Semantic Representation", *Cognitive Neuropsychology*, 2016, vol. 33(3–4), pp. 130–174.

[24] T. Litvinova and O.V. Dekhnich, "Modeling the Meaning of Individual Words Using Cultural Cartography and Keystroke Dynamics", *Integration of Education*, 2024, vol. 28(4), pp. 624 – 640.

[25] A. Miklashevsky, "Perceptual Experience Norms for 506 Russian Nouns: Modality Rating. Spatial Localization, Manipulability, Imageability and Other Variables", *Journal of Psycholinguistic Research,* 2018, vol. 47, pp. 641–661.

[26] P. Panicheva and T. Litvinova, "Matching LIWC with Russian Thesauri: An Exploratory Study", in Filchenkov A., Kauttonen J., Pivovarova L. (eds). *Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science*. Cham.: Springer; 2020, p. 181–195.

[27] T. Litvinova and P. Panicheva, "Individual Differences in the Associative Meaning of a Word Through the Lens of the Language Model and Semantic Differential", *Research Result. Theoretical and Applied Linguistics*, 2024, vol. 10(1), pp. 61–93.

[28] P. Seredin, D. Goloshchapov, V. Kashkarov, A. Emelyanova, N. Buylov, Y Ippolitov and T. Prutskij, "Development of a Visualisation Approach for Analysing Incipient and Clinically Unrecorded Enamel Fissure Caries Using Laser-Induced Contrast Imaging, MicroRaman Spectroscopy and Biomimetic Composites: A Pilot Study", *J Imaging*. 2022 May 13;8(5): 137.

[29] P. Seredin, D. Goloshchapov, Y. Peshkov, N. Buylov, Y. Ippolitov, V. Kashkarov, J. Vongsvivut and R. O. Freitas, "Identification of chemical transformations in enamel apatite during the development of fissure caries at the nanoscale by means of synchrotron infrared nanospectroscopy: A pilot study", *Nano-Structures & Nano-Objects*, vol. 38, 2024, 101205.